Computer Science Journal of Moldova, vol.**21**, no.3(63), 2013

Steering the Information Society Development as a Vision-based, Multi-stakeholder Process: a Point of View

Horatiu Dragomirescu, Florin Gheorghe Filip

Abstract

The steering of the information society development at country level is a desideratum for the smooth unfolding of the respective societal change – nowadays a key issue worldwide for scientific research and public governance. Some possible guidelines are presented on how to come up with a vision shared by stakeholders and how to ensure concertation among them along the way.

Keywords: Information Society, new and emerging IT, vision, stakeholders, concertation

1 Introduction

A few decades ago, Peter Drucker (1969) anticipated the "age of discontinuity", when foreseeing the future by extrapolating the facts from the past is very likely to be inefficient. He described four sources of discontinuities: a) new technologies, b) globalization, c) cultural plurality, and d) knowledge capital. Today's Information Society exhibits features that are attributable to above-mentioned factors.

The Information Society (sometimes called *e-Society*) is nowadays counted among those select issues that are top-ranked on the research agenda worldwide, as well as on the policy-making agenda of states and international organisations, such as the United Nations or the European Union. Interestingly, the tendency towards more and more diverse and refined ways of theorising the information society is paralleled by

the pragmatic quest for synergy in blueprints and initiatives within and across countries. Progressing towards the information society is a priority held in common by both developed and developing countries, although their respective approaches significantly differ. Developed countries generally focus on advancing R&D in the field of ICTs and the subsequent boost of market supply of more and more sophisticated goods and services. In turn, developing countries, faced with sharp *digital divide,* due to their lagging behind in terms of computer skills and infrastructures, are chiefly concerned with taking catching-up steps and obtaining international assistance.

The issue of information society development at large, including the particular aspect of its steering, calls for a pluralist stance being adopted in thinking and acting alike, in light of a consistent series of arguments:

- *theoretical*: the multitude of scientific disciplines concerned with studying the information society (sociology, anthropology, informatics etc.) the encounter of which turned the respective matter into a standalone interdisciplinary subject;

- *technological*: the interplay of several recent key and emerging technologies of the digital age (computing, telecommunications networks, broadband IP v6, wireless communication, cloud computing, RFID, social networks, Internet of things etc.) that came up to convergence and thus lead to new business and organisational models (pay-per-service, green computing, networked enterprise, smart factory) and types of human behaviour (virtual interactions, extended collaboration) (Bughin, Chui and Manyika, 2010; Zuehlke, 2008; Panetto, Jardin-Goncalves and Molina, 2012; Filip, 2013);

- *actional:* the development of the information society became an endeavour that is subject to enactment at various levels (national, regional, global) (Martin, 2005), a wide range of stakeholders being involved therein (the state, academia, business companies, the civil society, communities with professional, socio-cultural or territorial grounds, individual citizens).

The multiple ways in which the information society was defined along the timeline differ, depending of the perspective adopted, which can be either economical, social, technological, occupational or cultural (Webster, 2006, p. 8). Out of these, Kellerman (2002, p. 10) privileges the economical and the cultural ones; the information society is characterised, from the economic point of view, by the key role of information resources and information industries while, from the cultural one, it features a flourishing creativity and a rising intellectual standing of individuals.

At the second round of the World Summit on the Information Society, convened by the United Nations, in Tunis, in November 2005, there were consensually stated the "desire and commitment to build a people-centred, inclusive and development-oriented Information Society" (WSIS, 2005). Such an option is in line with the bolder emphasis on "I" term (information), within the "IT" construct (Davenport, 2000), and also with the reinforced anthropocentric and collaborative orientation in designing next-generation information systems. It is also in line with "VISION 2050: the New Agenda for Business" of the World Business Council for Sustainable Development, released at the World CEO Forum in New Delhi, in 2010 (WBCSD, 2010).

2 The path of the information society development

The information society development at country level can be understood as a process of evolving towards upper maturity stages of this type of societal system; the knowledge society represents the fully mature stage attainable by an information society over time.

At global scale, as Himanen (2004) remarked, "[T]he first phase of the information society focused on the development of technology, such as network connections. In the second phase, which has now begun, technological development will continue; however, the focus will shift to larger social matters and the main focus will be on changing the ways in which we operate."

The information society development path at country level can be

mapped in several modalities. The qualitative ones are generally based upon scales of successive development stages, considered either prospectively or in retrospect. Each stage is assigned specific features the actual presence of which reveals the attainment of a certain maturity degree, with a view to assessing progresses achieved or projecting the way ahead.

Miles (2002, p. 163) proposed a range of maturity stages defined in metaphorical terms: islands, archipelagos, continents, ecosystem; this approach is useful in that it renders the bottom-up information society development path intelligible, although in rather intuitive terms. Accordingly, its steering process is meant to foster the consolidation of punctual instantiations into an integrated whole.

A scale with a more analytical format was proposed by Rao (2005), in which the informational society development stages range from the incipient ones (disarticulated, embryonic), followed by those where maturation is still underway (development, concertation, intermediate), up to the top ones (mature, advanced, world leader); the respective stages are distinguishable against 8 characteristics (connectivity, content, communities, commerce, culture, capacities, cooperation, capital).

Quantitative approaches, in turn, involve the use of specific metrics aimed at visualising the development degree reached by the information society at different levels (country, region, multi-country). Dedicated composite indexes became more popular lately and are widely used, especially by international bodies; the reasons are their transparent computation methodologies and the easiness of their assessment by policy-makers, as well as by the public at large. Their shortcomings include the fact that they are merging several dimensions, each of them having a standalone significance that could be disguised through statistical consolidation; moreover, if countries in upper ranks are perceived as de-facto standards, a race for higher scores is triggered to the detriment of cooperation. For monitoring purposes, specific indicators can also be used in their elementary forms, but they become more informative if appropriately selected and assembled into meaningful scorecards that allow for synoptic grasp.

3 The need for shared vision and concertation among stakeholders

Although the passage from the industrial society to the information society was recognised as a global mega-trend, this shift does not take place by default; an overall steering is required to ensure its pertinent orientation, sustained pace and beneficial systemic impact. Such a steering should be exerted as a vision-based, multi-stakeholder process.

The importance of adopting a multi-stakeholder approach was stressed, among others, by UNESCO; under the aegis of its "Information for All" programme, a template was issued for designing national information policy and strategy frameworks pertaining to the development of the Information Society at country level (Finquelievich, 2009).

According to Afsarmanesh and Msanjila (2010, p. 62), "[A] vision is a deeply held picture of where a person, a group of people, an organization, or a society, wants to reach in the future." It is meant to lay down the main directions towards meeting the final goal, without setting specific targets and deadlines. Foresight studies are always useful in preparing such visions of information society development at country level (Filip *et al.*, 2004; Dragomirescu and Filip, 2008)

In Europe, Finland provides an insightful example of good practice in steering the Information Society development at country level; as mentioned by Repo (2003), the Finnish Information Society Development Centre (TIEKE) "plays a key networking role in connecting various players in the development of the Finnish Information Society."

Yet another relevant example refers to Ireland, where an Information Society Steering Committee was established, in 1996, by the Minister for Employment and Enterprise, followed by the creation of the Inter-departmental Implementation Group on the Information Society (IIGIS); the government's strategy for the information society was originally laid down, in March, 2003, in the document titled "New Connections: A Strategy to realise the potential of the Information Society", progress reports having been published regularly (Mc Caffery, 2007).

Understandably, the state always plays a key role, as it is in charge with designing and implementing public policies, also administering

public spending; besides, the public sector is a major information producer. However, especially in developing countries, state bodies primacy in steering the information society development can be perceived by the other stakeholders as hegemonic, reverse effects being thus likely to occur as pluralism is weakened. A viable alternative in this respect consists of the lead being taken by independent, highly representative and forward-thinking bodies such as countries' academies of science (Dragomirescu and Filip, 2008). Consequently, governmental agencies set up to supervise the development of the information society should rather focus on operational aspects, acting on behalf of the state as one of the several stakeholders concerned.

The shared vision, once validated, allows for harmonising the unfolding of individual stakeholders' self-managed initiatives as well as their participation in joint actions. Concertation among stakeholders is a challenging endeavour though, as they have different logics and also maintain their autonomy. Cultivating autonomy in terms of choosing means and practical solutions is, nevertheless, fully compatible with cooperation on the grounds of the shared vision on information society development. At the same time, the stakeholders should be aware that e-Society might create new problems. Some of them can be seen in Table 1.

In order for stakeholders to keep a sustainable engagement, it is essential that the respective vision be generated collaboratively; further on, a regular consultation framework should be established for updating and progress evaluation purposes.

4 Conclusions

Steering the development of the information society at country level involves balancing top-down and bottom-up approaches, learning from experience and envisioning, incremental and leapfrog types of dynamics, and wisely matching means to the goals pursued. Adopting a systemic mindset, a collaborative format of the relationship among stakeholders, and smart methods of consultation and coordination are among the key success factors in performing the steering as a win-win enterprise.

Table 1. e-Society: pros and cons

Advantages	**Debatable impacts**
• More effective time usage • Facilitation of communication between people placed in different contexts • More comfortable work conditions • Good ecological impact	• Uniformization of human behaviour • Limiting direct human contacts • Laziness tendencies • High risks of manipulation • Functional opacity • High vulnerability of private

Source: Filip (2013)

Acknowledgement

An earlier version of this paper was published in C. Gaindric and S. Cojocaru (eds.), *Proceedings of the International Conference on Intelligent Information Systems*, IIS 2013, August 20 - 23, Chisinau, Republic of Moldova, Institute of Mathematics and Computer Science, Chisinau, pp. 5–9.

References

[1] H. Afsarmanesh, S. Msanjila. (2010). *ePAL 2020 Vision for Active Ageing of Senior Professionals.* In: Camarinha-Matos, L. M., Boucher, X. and Afsarmanesh, H. (eds.), "Collaborative Networks for a Sustainable World". Proceedings

of the 11^{th} IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2010, St. Etienne, France, October 11-13, 2010, Springer, Boston, pp. 60–72.

[2] J. Bughin, M. Chui, J. Manyika. (2010). *Clouds, big data and smart assets: ten tech-enabled business trends to watch.* McKinsey Quarterly, Iss. 4, pp. 26–43.

[3] T. Davenport. (2000). *Putting the I in IT.* In: D. Marchand and T. Davenport (eds.), Mastering Information Management, Pearson Education, London, pp. 5–9.

[4] H. Dragomirescu, F. G. Filip. (2008). *The Knowledge Society Agenda in Romania: from Experts' Vision to Public Perception.* In: M. D. Lytras et al. (eds.), "The Open Knowledge Society. A Computer Science and Information Systems Manifesto". Proceedings of the 1st World Summit for the Knowledge Society. Communications in Computer and Information Science, vol. 19, Springer, Berlin, Heidelberg, pp. 24–31.

[5] P. Drucker. (1969). *The Age of Discontinuity.* Harper & Row, New York.

[6] F. G. Filip, H. Dragomirescu, R. Predescu, R. Ilie. (2004). *IT tools for foresight studies*, Studies in Informatics and Control, Vol. 13, No. 4, pp. 161–168.

[7] F. G. Filip. (2013). *Sustainable development and the emergent Information and Communication Technologies.* Keynote speech, 3^{rd} International Conference on Logistics, Informatics and Service Sciences (LISS 2013), Reading, UK, 21 - 24 August, 2013. Available at: http://www.icir.bjtu.edu.cn/liss2013/keynote_speackers.asp (accessed 21.09.2013).

[8] S. Finquelievich. (coordinator) (2009). *National Information Society Policy: A Template.* UNESCO, Paris. Available at: http://unesdoc.unesco.org/images/0018/001871/187135e.pdf (accessed 08.07.2013).

[9] P. Himanen. (2004). *Challenges of the Global Information Society. Committee for the Future,* Parliament of Finland, Helsinki. Available at: http://web.eduskunta.fi/dman/Document.phx?documentId= br11307103930385&cmd=download (accessed 08.07.2013).

[10] A. Kelerman. (2002). *The Internet on Earth: A Geography of Information*, Wiley, London and New York.

[11] B. Martin. (2005). *Information Society Revisited: from Vision to Reality.* Journal of Information Science, Vol. 31, No. 1, pp. 4–12.

[12] C. Mc Caffery (2007). *Irish Information Society Policy.* Networks and Communication Studies (NETCOM), Vol. 21, No. 1 - 2, pp. 209–240.

[13] I. Miles, (2002). *Information Society Revisited: PICTuring the Information Society.* In: R. Mansell, R, Samarajiva and A. Mahan (eds.), "Networking Knowledge for Information Societies: Institutions & Intervention", Delft University Press, Delft, pp. 160–165.

[14] H. Panetto, R. Jardin-Goncalves, A. Molina. (2012). *Enterprise integration and networking: Theory and practice.* Annual Reviews in Control, Vol. 36, No. 2, pp. 284–290.

[15] M. Rao. (2005). *The Information Society: Visions and Realities in Developing Countries.* In: O. Hemer and T. Tufte (eds.), "Media & Global Change. Rethinking Communication for Development", CLACSO, Consejo Latinoamericano de Ciencias Sociales, Buenos Aires, pp. 271–284.

[16] A. Repo. (2003). *TIEKE – National Meeting Point for Information Society Developers.* The Innovation Journal, Vol. 8, No. 4. Available at: http://www.innovation.cc/discussion-papers/repotools.pdf (accessed 21.09.2013).

[17] WBCSD (2010). *Vision 2050: The new agenda for business,* World Business Council for Sustainable Development secretariat, Geneva. Available at: http://www.wbcsd.org/Pages/Adm/Download.aspx?ID=320& ObjectTypeId=7 (accessed 21.09.2013).

[18] F. Webster. (2006). *Theories of Information Society (third edition),* Routledge, Abingdon and New York.

[19] WSIS (2005). *Tunis Commitment.* Document WSIS-05/TUNIS/DOC/7–E, 18 November, 2005. Available at: http://www.itu.int/wsis/docs2/tunis/off/7.html (accessed 08.07.2013).

[20] D. Zuehlke. (2008). *SmartFactory – from vision to reality in factory technologies.* In: M. J. Chung and P. Misra (eds.), Proceedings of the 17th International Federation of Automatic Control (IFAC) World Congress (plenary paper), Seoul, 6 – 11 July, 2008, pp. 14101–14108. Available at: http://www.nt.ntnu.no/users/skoge/prost/proceedings/ifac2008/data/papers/4283 (accessed 21.09.2013).

Horatiu Dragomirescu, Florin Gheorghe Filip Received September 25, 2013

Horatiu Dragomirescu
Bucharest University of Economic Studies – ASE, Piata Romana 6,
Bucharest, Romania
E–mail: *dragomir@ase.ro*

Florin Gheorghe Filip
Romanian Academy, BAR & INCE, Calea Victoriei 125,
Bucharest, Romania

ICI Bd. Averescu 8-10, Bucharest, Romania
E–mail: *ffilip@acad.ro*

Computer Science Journal of Moldova, vol.**21**, no.3(63), 2013

New Vision and Goals of Informatics and Megachallenges of Mankind. Extended abstract

Jozef Gruska

1 A birth of modern Informatics

Currently dominating perception of computer science has its origin in a very cleverly written, and much influential, paper of *Newel, Simon and Perlis*, published in Science in 1967, that well captured the perception of the field at that time.

The basic ideas presented in their paper were:

" *Whenever there are phenomena there can be a science dealing with these phenomena. Phenomena breed sciences. Since there are computers, there is computer science. The phenomena surrounding computers are varied, complex and rich.*"

Since that time there have been numerous attempts to modernize such a view of computer science. Some of them centered around a slightly modified name of the field as *computing science* with attempts to put emphases on algorithms, programming and software instead of hardware.

Some of the first along these lines was Dijkstra with his position: *"Computing science is - and will always be - concerned with the interplay between mechanized and human symbol manipulation usually referred to as "computing", and "programming", respectively. It is located in the direction of formal mathematics and applied logic, but ultimately, far beyond where those are now."* This was underlined by his famous, satiric but deep vision that *Computer science is as much about computers as astronomy is about telescopes.*

In spite of all these attempts, computer-centric or computing-centric view of computer science still dominates.

One of the first areas of Informatics that abandon such a technology dominating view of the field was artificial intelligence that started (and had to start) to divorce with the main stream of computer science, fortunately.

One example to demonstrate how deep was such computer-centric view of the field can be found in the development of IFIP (International Federation for Information Processing), that played, especially its world congresses, technical committees and working groups, such an important role in the development of the field. Though IFIP was established in 1962, till 1996 there was no technical committee for theory and the reasoning behind was that the only theory behind information processing is theory connected with programming and the rest of "theory" belongs either to mathematics or to electrical engineering.

There are nowadays a variety of reasons why such a computer-centric view of the field should be seen as obsolete, not broad and not deep enough, and actually damaging the development of the field. They will be discussed only briefly in this paper, for more see [1]. Here are some of the reasons.

- An understanding starts to be developed that information processing plays key role both in physical and biological nature. For example, quantum, DNA and molecular information processing. In particular, an understanding developed that information processing is of such an importance for life as breathing and eating and that even very primitive living being can do exceptionally complex and efficient information processing.

- All natural sciences, and not only these sciences, are starting to be increasingly seen as being, to a large extent, information processing driven, and not only that. It starts to be understood that all sciences start to converge to Informatics once seen in a proper broadness and deepness.

- On a more practical level, it starts to be clear that in the coming future any very significant innovation will use advanced Informatics tools, methods and paradigms.

All that requires that a much broader and deeper view of the field should be developed – see [1, 2].

2 A new perception of Informatics

A new perception of the Informatics here presented sees the field as consisting of four much interleaved components:

- scientific Informatics;
- technological Informatics;
- new methodology;
- applied Informatics.

As a scientific discipline of a very broad scope and deep nature, Informatics has many goals. Its main task is to discover, explore and exploit in depth, the laws, limitations, paradigms, concepts, models, theories, phenomena, structures and processes of both natural and virtual information processing worlds.

To achieve its tasks, scientific Informatics concentrates on new, information processing based, understanding of universe, evolution, nature, life (both natural and artificial), brain and mind processes, intelligence, creativity, information storing, processing and transmission systems and tools, complexity, security, and other basic phenomena of information processing worlds.

Development and analysis of a variety of formal, descriptional, computation, interaction and communication models and modes, development and analysis of (deterministic, randomized, genetic, evolutionary, quantum, ...) algorithms, protocols and games are some of the main tools of Informatics.

Data, information, knowledge, formal systems, logics, algorithms, protocols, games, resources, models and modes of information processing, communication and interactions are the key concepts behind.

In order to meet its goals, Informatics develops close relations with other sciences and technology fields, especially with physics and biology, on the one hand, and with electronics and nanotechnologies on the other.

Informatics as a science includes also numerous theories much needed for its development to depth and in broadness. Some theories are very abstract, others quite specific, and some theories are oriented on making better use of the outcomes of the scientific Informatics to create a scientific basis of Informatics as of an engineering/technology discipline.

One way to illustrate such a broad and deep perception of scientific Informatics will be in this paper through presentation and analysis of its grand challenges. This will be discussed briefly below. In the same way one can illustrate main tasks of technological and applied Informatics, but this is beyond the scope of this paper, see [1, 2].

Another way to illustrate such a broad and deep view of scientific Informatics is to make an analogy between views of Physics and Informatics because their goals can be seen as being very similar.

The main goal of *Physics* can be seen as to study laws, limitations and phenomena of the *physical worlds*.

The main goal of *Informatics* can be seen as to study laws, limitations and phenomena of the *information worlds*.

Physics and Informatics can therefore be seen as representing two windows through which we try to perceive and understand the world around us.[1]

2.1 Grand challenges of scientific Informatics

New main grand challenges of scientific Informatics can be briefly summarized as follows:

[1] In a similar way we can see life-sciences and Informatics as providing two windows and tools with which we try to understand, imitate and outperform the biological world and its highlights – human brain, mind, consciousness, and cognitive capabilities.

- To explore our world as a point in a space of potential information processing worlds.
- To explore laws and limitations of information processing that governs universe, evolution and life.
- To develop theoretical foundations for design, analysis, verification, security, simulation and modeling of huge information processing systems.
- To understand intelligence, creativity, mind and consciousness.
- To make foundations for science and engineering of the science making activities.
- To understand and manage all aspects of computation, communication and structural complexity.

3 Informatics-driven methodology

Of a key importance for a new perception of Informatics is also an understanding that Informatics, as a symbiosis of a scientific and a technology discipline, develops also basic ingredients of a new, in addition to theory and experiments, the third basic methodology for all sciences, technologies and society in general.

This new, Informatics-based, methodology provides a new way of thinking and a new language for sciences and technologies, extending the Galilean mathematics-based approach to new heights.

The main components of this new methodology can be briefly summarized as follows:

- Simulation methods and systems.
- Modeling – Design of information processing models.
- Visualisation and animation.

- Searching (sophisticated searching as an alternative to deep knowledge based reasoning).
- Design of systems with superhuman intelligence.
- Design of systems for problem solving and reasoning.
- Development of methods to specify, design, analyse, verify and reliably run complex (information processing) systems.
- Design of algorithms, study of their performances and study of inherent complexities of computational, communication and description systems.
- Design, analysis and comparison of descriptional languages and systems and of the relations between objects and their specifications.
- To study problems of the real world as of the one of information processing worlds.

Informatics-driven methodology subsumes and extends the role and improves tools Mathematics used to play in advising, guiding and serving other scientific and technology disciplines and society in general.

Power of the new, Informatics driven methodology, is discussed in the paper [1] in details. Here are only few of the main reasons:

- New methodology brings new dimension to both old methodologies.
- It brings into new heights an enormous power of modeling, simulations and visualisation.
- It utilises an enormous exploratory and discovery power of automata, algorithms and complexity considerations.
- It utilizes enormous exploratory power of the development and study of artificial, men made systems, for understanding of phenomena of natural phenomena and systems.

- It utilizes enormous discovery and exploratory power of the correctness and truth searching considerations, systems and tools.
- It utilizes an enormous potential that the study of virtual worlds brings for an understanding of the real worlds.
- It seems to have a big chance to make hard sciences from (at least some) of the soft sciences.

4 Informatics and new megachallenges of science and technology

Because of its enormous guiding power for practically all areas of science, technology and the whole society and an enormously powerful tools Informatics offers, we can see Informatics as a new queen and at the same time a new powerful servant for all of society.

In particular Informatics is expected to play the key role in dealing with two main megachallenges of current science, technology and society. Namely:

- **To beat natural human intelligence.** More exactly, to create super-powerful non-biological intelligence and its merge with biological intelligence.
- **To beat natural human death**. More exactly, to increase much longevity for human bodies and to achieve uploading for human minds. In more details, to fight natural death as another disease and to find ways to upload human mind to non-biological substrate.

Mankind starts to have enough reasons to see the above megachallenges as being currently realistic enough. Here are some of them.

- Because computers performance keeps developing faster and faster, actually exponentially, there are good reasons to assume

to have soon (around 2045?) even laptops with information processing power and capacity larger than of all human brains – see [3].

- Exponential scaling up concerns also of all main information related technologies, especially genetic and nanotechnologies as well as artificial intelligence. This creates another basis for making two megachallenges as already feasible ones.
- Exponential developments of information processing technologies are believed to imply enormous speed up in developments of main sciences and technologies.
- Tools to reverse engineering brains keep also developing exponentially concerning their potential and precision and so we can assume to have quite soon ways to simulate functionality of human brains.
- Society keeps putting enormous effort, actually more and more human and money resources, to develop and apply genome engineering, to model human brains and minds as well as to vastly extend human longevity.
- A vision starts to be accepted to see the development of superintelligent machines as the next stage of evolution and to prepare society for handling and accepting such developments.

To deal with new megachallenges practically all areas of sciences and technologies have to be involved. However, Informatics is expected to play by that a very important role for several reasons.

- It starts to be clear that in order to understand more deeply functionality of living systems, from cells to brains, and to design using other, non-biological substrates, systems to outperform them, information processing models of such systems are needed. Chemistry and biology has been able to gather enormous number of data about composition and behaviour of elements of particular living systems, but Informatics tools are needed to model their

functionality as complex systems in such a way that we can model their functionality using non-biological substrates. It starts to be understood that modeling on the basis of differential equations can hardly lead to design of efficient models and that modeling using Informatics tools to model concurrent and parallel systems may be needed.

- New, Informatics-driven methodology, is expected to increase also exponentially the whole development of science and technology, in all their areas, and this is another essential reasons that we can expect to achieve in few decades what seems to us as needed several hundred years.

5 Food for thoughts

New megachallenges for science and technology, and actually for the whole society, to which Informatics is to contribute essentially, aim to bring enormous, hard to imagine to almost anyone few years ago, changes to the lives, value systems and goals of individuals as well as to the life and goals of the whole society. In spite of all revolutionarity of these goals, some of the best minds of society started to foresee them immediately when enormous potential of modern computers started to be revealed. In the same way, deep thinkers have also already formulated in a short but sharp way the essential problems we are to encounter and even the way to deal with them.

In the following some of these deep thoughts, worth to notice and rethink, are presented.

- There is plenty of room at the bottom. *R.Feynman addressing American Physical Society in 1960 – seen nowadays as first understanding of the potential of nanotechnologies.*
- There is nothing in biology found yet that indicates the inevitability of death. *Richard Feynman*
- It seems probable that once the machine thinking method had started, it will not take long to outstrip our feeble power. They

would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to expect machine to take control. *Alan M. Turing, 1952*

- The ever-accelerating progress of technology .... gives the appearance of approaching some essential singularity in the history of the [human] race beyond which human affairs, as we know them [today] could not continue... *John von Neumann, 1950.*

- Let an ultraintelligent machine be defined as a machine that can far surpass all intellectual activities of any man, however clever. Since the design of machines is one of intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be and "intelligent explosion" and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man needs ever make. *I. J. Good, 1965, a British mathematician.*

- Since there is a real danger that computers will develop intelligence and take over we urgently need to develop direct connections to brains so that computers can add to human intelligence rather than be in opposition. *Stephen Hawking, 2011*

One can even go more back to the history to see that seeds of the vision of the future have already appeared long time ago once machines started to be used to improve power of mankind. Indeed, Samuel Butter, an English writer, wrote in 1863, 4 years after the publication of Darwin's *The origin of species* the following thoughts:

There are few things of which the present generation is more justly proud than the wonderful improvements which are daily taking place in all sorts of mechanical appliances.....

But what would happen if technology will continue to evolve so much more rapidly than the vegetable and animal kingdom? Would it displace us with supremacy of earth?

Just as the vegetable kingdom was slowly developed from the mineral one and, similarly, the animal kingdom from the vegetable, so in these

last few ages an entirely new kingdom has sprung up, of which as yet we have only seen what will one day be considered as a prototype of the race....

We are daily giving [machines] greater power and supplying them by all sorts of ingenious contrivances, such as self-regulating and self-acting power, which will be to them what intellect has been to human race.

References

[1] Jozef Gruska. *A perception of Informatics*, web page of Academia Europaea, http://www.AE-Info.org/ae/user/Gruska.Jozef, 55 pages.

[2] Jozef Gruska. *Impulses and roads to a new perception of informatics*, in "Rainbow of Computer Science", edited by C. Calude, A. Salomaa and G. Rozenberg, Springer Verlag, 2011, pp. 183–199.

[3] Ray Kurzweil. *Singularity is near*, Penguin books, 2007.

Jozef Gruska

Faculty of Informatics, Masaryk University
Brno, Czech Republic
E–mail: *gruska@fi.muni.cz*

Received October 3, 2013

Computer Science Journal of Moldova, vol.**21**, no.3(63), 2013

Increasing the Effectiveness of the Romanian Wordnet in NLP Applications*

Verginica Barbu Mititelu

Abstract

The Romanian wordnet is a semantic network under ceaseless enrichment and improvement. Its use in various applications throughout time highlighted the need for further development. In this paper we focus on a question answering scenario. We show how adding derivational relations between the literals already present in the network could help increase the effectiveness of using the Romanian wordnet in such an application. We describe the steps we took in the process of identifying, validating and adding derivational relations in our network and then simulate a question answering situation using RoWikipedia as corpus.

Keywords: wordnet, Romanian, derivational relations, question answering, lexical chains.

1 Introduction

Applications in the Natural Language Processing (NLP) domain need quality language resources for attaining good results. These resources can be lexicons, dictionaries, thesauri, grammars, etc. In this article we focus on the knowledge about words and their meanings, on the way it is represented so that to facilitate its effective use in NLP applications.

Among the various formalisms available for representing lexical knowledge, semantic networks are the most widely known and used. Furthermore, a wordnet is the most popular kind of semantic network.

*This work was supported by the Sectorial Operational Program Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number SOP HRD/89/1.5/S/59758.

It only contains nouns, verbs, adjectives and adverbs, as they make up the lexical component of a language; prepositions and conjunctions belong to the syntactic component [13], fulfilling a relational function.

In this net words are organized according to psycholinguistic principles, by means of semantic relations, many of them specific to certain parts of speech. Thus, hyponymy and meronymy are specific to nouns; hyponymy, troponymy, lexical entailment and cause are proper for verbs; descriptive adjectives are organized in clusters based on their similarity of meaning; relational adjectives are linked to the corresponding nouns, while adverbs are linked to the respective adjectives.

The first such language resource created was the Princeton WordNet (PWN henceforth) [14, 4]. Since 1985 it has been under quantitative and qualitative improvement. It served as a model for similar resources for tens of other languages. In 1996, within the EuroWordNet project [19], semantic networks started being developed for 8 European languages (Dutch, Spanish, Italian, English, French, German, Czech, and Estonian) after the model of PWN. In 2001, within the BalkaNet project [18] wordnets for Bulgarian, Czech, Greek, Romanian, Serbian and Turkish began being created (or continued being developed in the case of Czech). Nowadays there are well beyond 60 languages for which such a resource was created or is being created (a list is available here: `http://globalwordnet.org/gwa/wordnet_table.html`). The developers and users constitute a very active community, holding their conference (Global WordNet Conference) every two years in various locations around the world, Asia being a frequent host of the meetings. Moreover, all major conferences in the field of Natural Language Processing and Computational Linguistics accept papers on wordnets.

Given the success such resources have among researchers, linguistically, one can notice that the term "WordNet" has become a class name, so a common noun, and is used in the form "wordnet" to refer to any semantic network realized after the model of PWN [5].

The Romanian wordnet (RoWN henceforth) has been being developed since 2001. During BalkaNet, a common team from the Romanian Academy Research Institute for Artificial Intelligence and from the Faculty of Informatics of the "Al. I. Cuza" University of Iaşi worked for

developing a core of 18000 synsets, conceptually aligned to PWN and through it to the synsets of all the BalkaNet wordnets. The concepts considered highly relevant for the Balkan languages [18] were identified and implemented first, then a set of concepts specific to the Balkan area. After the BalkaNet project ended, the Romanian Academy Research Institute for Artificial Intelligence undertook the task of maintaining and further developing the RoWN.

We selected the concepts to be further implemented in our network so that they served other tasks that we accomplished throughout time. Thus, we aimed at a complete coverage of the 1984 corpus (`http://nl.ijs.si/ME/Vault/CD/docs/1984.html`), of the newspaper articles corpus NAACL2003 (possible to be searched for at `http://ws.racai.ro:9191`), of the Acquis Communautaire corpus (`http://ipsc.jrc.ec.europa.eu/index.php?id=198`) and of the Eurovoc thesaurus (`http://eurovoc.europa.eu/`), as much as possible from the Wikipedia lexical stock, and the verbs in VerbNet (`http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`).

We continued to follow the methodology established during the BalkaNet project, following the expand model [17]. Two basic development principles have always been followed: the Hierarchy Preservation Principle (according to which the hierarchical structure of the concepts in a wordnet is the same irrespective of the natural language for which the wordnet is developed) and the Conceptual Density Principle (which ensures that once a concept is selected to be implemented, all its ancestors up to the unique beginners are also selected, thus preventing the existence of dangling nodes) [18].

2 NLP Applications and RoWN

The ceaseless development of RoWN is (also) justified by its use in various applications implemented in our Institute. We enumerate below these applications and the way RoWN served their aims.

- Word Sense Disambiguation (WSD):

  - In a monolingual context [8]: lexical chains between the various senses of the words in the sentence are looked for in the RoWN and, when found, their length is calculated by counting the number of nodes and of edges crossed to get from one end of the chain to the other: the shorter the lexical chain, the smaller the length, so the more related the linked words senses; obviously, the lexical chain is shorter when there are more relations in the network;
  - In a multilingual context [6]: conceptually aligned wordnets for various languages permit disambiguation of homographs in one language due to their translation by different words in (an)other language(s); the results in this case were reported as better than those of WSD in a monolingual context.

- Question Answering (QA) [7]:
  - In order to automatically find the answer to a user's question formulated in natural language, the system relies only on the words introduced by the user. However, these are not always the best chosen ones. (Imagine the trivial case of non-native speakers of a language looking for information in that respective language.) That is why, it can be necessary to use also synonyms, hypernyms, hyponyms, troponyms or derived words from the ones introduced by the user. Due to its organization, a wordnet can offer access to these words for expanding the user's query, so that the sentences containing the answer could be found more easily and more reliably.
  - For ordering the answers found by the system according to their relevance in respect to the user's question, it is necessary to find a semantic similarity score between the words introduced by the user and the words occurring in the text (as they may not be the same); for calculating this score the length of the lexical chains between the respective words in the wordnet is considered. The shorter the chain, the more similar the question and the found text, so the higher the probability for it to be the answer to the user's question.

In a QA task in a multilingual context (i.e., the user asks a question in one language and needs to find the answer in texts written in a different language), conceptually aligned wordnets prove their usefulness for the cross-lingual equivalence of terms (see a detailed description in [1]).

- Machine translation: conceptually aligned wordnets for more languages are a source of equivalent words and (simple and multiword) terms useful for feeding a translation table.

3 Adding Value to the RoWN by Marking Derivational Relations

RoWN has been developed by following the expand method and obeying the Hierarchy Preservation Principle (and the Conceptual Density Principle, see above). Thus, the semantic relations in PWN have been transferred into RoWN and organize its content, too. At the moment, the distribution of synsets and literals in RoWN is indicated in Table 1.

Table 1. Statistics about RoWN – synsets and literals

Part of Speech	Synsets	Literals	Unique Literals	Non-lexicalized
Nouns	41063	56532	52009	1839
Verbs	10397	16484	14210	759
Adjectives	4822	8203	7407	79
Adverbs	3066	4019	3248	110
TOTAL	59348	85238	75656	2787

As far as semantic relations are concerned, their occurrence in our RoWN is presented in Table 2.

With the exception of "attribute" relation, all the others enumerated in Table 2 link synsets with literals of the same part of speech. A path between two words of a different part of speech, about which any speaker would say they are related, although not impossible to

Table 2. Relations in RoWN

Relation	Number
hypo/hyperonymy	48316
instance_hypo/hyperonymy	3889
antonym	4131
similar_to	4838
verb_group	1530
member_holonym	2047
part_holonym	5573
substance_holonym	410
also_see	1333
attribute	958
cause	196
entailment	371

find, would be too long, thus providing wrong information about the similarity between those words.

Besides semantic relation, PWN also contains lexical relations, which are established between literals, unlike semantic ones which hold between synsets. Lexical relations are synonymy, antonymy, derivational relations. Involving literals, they are language specific, so cannot be transferred cross-lingually. It is worth noticing in Table 2 that antonymy, which is a lexical relation in PWN, is represented as a semantic one in RoWN. The conceptual opposition between the synsets containing the antonymic pair is more useful in various applications than the mere antonymy between two literals, that is why we extended the antonymy relation from PWN at the synsets level in RoWN.

PWN also contains derivational relations. Although many of them have a correspondent in Romanian, they cannot be automatically transferred into RoWN. Such a strategy of enriching wordnets with derivational relations does exist in the wordnet community [10, 11, 12]. However, we preferred to find a language internal strategy for identifying

derivationally related words in our language and for marking them in RoWN (others who report similar attempts are [3, 15, 16, 9]). Examples of cases when there is a derivational relation in PWN but no corresponding one in RoWN between literals lexicalizing the same concepts include: *prick* - *pricker* (Romanian: *înţepa* - *sulă*), *pacify* - *pacifier* (Romanian: *împăca* - *suzetă*), *dip* - *dipper* (Romanian: *afunda* - *polonic*), etc.

In order to mark such relations in our RoWN, we followed the steps below:

1. Find possible pairs of root-derived words among the (31872) simple literals in RoWN using a list of (492) Romanian affixes and then validate the pairs. We searched for pairs of literals (literal_1 and literal_2) such that literal_1 +/− affix(es) = literal_2. The "+" version covers progressive derivation, while the "−" version covers backformation. We allow for at most 2 affixes, but of different types. The results are in Table 3.

Table 3. Derived words in RoWN

Derivation type	Derived words	Percent
Prefixation	2862	17.43
Suffixation	13556	82.57
TOTAL	16418	

We subject the found pairs to an automatic validation and then to a manual one. For the former, we relied on the information about the part of speech of the words to which affixes can attach and of the words they help create. For example, the suffix *-a* can be attached to nouns or to adjectives to create verbs.

Afterwards we proceeded to a manual validation of the whole number of pairs. The results are presented in Table 4: for each type of derivation (prefixation or suffixation), from the found pairs (column 2) we present the number of those passing the automatic validation in column 3 and then of those that passed

the manual validation in column 4; the last column presents the percent of validated pairs for each derivation type.

Table 4. Evaluation of derived words from RoWN

Derivation type	Found	Automatic Validation	Manual Validation	%
Prefixation	2862	2621	1990	69.53
Suffixation	13556	8345	8452	62.35
TOTAL	16418	10966	10442	-

2. Extract (in a set) all synsets in which each member of the above validated pairs occurs; calculate the Cartesian product of the sets for a pair of literals; validate the members of the Cartesian product, thus obtaining a list of pairs of word senses between which a derivational relation was marked (notice that it is not valid at the synset level, but at the literal one). The results are in Table 5.

Table 5. Annotated pairs in RoWN

	Prefixed	Suffixed	TOTAL
Pairs subject to validation	30132	25717	55849
Validated pairs	3145	13916	17061
Percent	10.43	89.64	30.55

3. Add a semantic label for each derivational relation in the form of a semantic relation in the network between the synsets to which the literals in derivational relation belong. A statistics of these labels can be found in [2].

Marking such relations in our wordnet, we increased the number of cross-part of speech relations to a high extent, as 66% of the suffixed

words and 97% of the prefixed words have a different part of speech from their root.

4 Short Demonstration

For proving that adding derivational relations to the RoWN we increase its effectiveness in NLP applications, let us consider the QA task. Our corpus for searching answers can be RoWikipedia. One possible question of a user is "Cine a inventat motorul cu reacţie?" ("Who invented the jet engine?"). A sentence such as "Henri Coandă a inventat motorul cu reacţie." ("Henri Coandă invented the jet engine.") does not occur in RoWikipedia. However, one can find the answer in the corpus sentence "Henri Marie Coandă (n. 7 iunie 1886 - d. 25 noiembrie 1972) a fost un academician şi inginer român, pionier al aviaţiei, fizician, inventator, inventator al motorului cu reacţie şi descoperitor al efectului care îi poartă numele." ("Henri Marie Coandă (born 7 June 1886 - died 25 November 1972) was a Romanian academician and engineer, pioneer of aviation, physicist, inventor, inventor of the jet engine and discoverer of the effect bearing his name."). The only term common to both the question and the answer is "motor cu reacţie" ("jet engine"). This unique match is not enough for giving a high score to the sentence so that it should be returned to the user. However, expanding the query, the system will also search for words that are semantically related to those introduced by the user. So, one more match will be possible: between "inventat" and "inventator". In fact, the maximum number of matches is now complete, so the sentence is retained by the system.

For calculating the semantic distance or similarity between two word senses lexical chains are created, i.e., the links and nodes in the network that are crossed for getting from one node (containing one of the target word sense) into another (containing the other target word sense). The shorter the chain, the more similar the senses. For the pair "inventa" (occurring in the user's question) - "inventator" (occurring in the corpus), the lexical chain between them crossed 6 nodes and 7 relations previously:

inventator(1.1) *instance_hyponym* James_Watt(x)
James_Watt(x) *instance_hypernym* inginer(1.1)
inginer(1.1) *hyponym* inginer_software(1)
inginer_software(1) *domain_member_TOPIC* ştiinţa_calculatoarelor(x)
ştiinţa_calculatoarelor(x) *domain_TOPIC* programa(3)
programa(3) *hyponym* crea_mental(1)
crea_mental(1) *hypernym* inventa(1)

The strangeness of this example results from the intricate path from *inventator* to *inventa*, uncommon for whatever speaker of Romanian: *inventator* - *James Watt* - *inginer* "engineer" - *inginer software* "software engineer" - *ştiinţa_calculatoarelor* "computer science" - *programa* "to program" - *crea_mental* "to create by mental act" - *inventa*. Now that derivational relations are marked, there is a direct link (semantically labeled *agent*) between the two words:

inventator(1.1) *agent* inventa(1).

5 Conclusions

Derivational relations need to be marked in a wordnet due to several reasons: derived words are part of our mental lexicon (although speakers also know the rule for creating derived words) and are in semantic relations to their roots, creating micro-networks. Moreover, from a practical perspective, the more relations are marked in the wordnet, the more effective it becomes in the applications it is used in. We have proved this in a QA scenario for Romanian. A rerun of the QA algorithm working with the enriched RoWN must support our demonstration.

References

[1] V. Barbu Mititelu, Alexandru Ceauşu, Radu Ion, Elena Irimia, Dan Ştefănescu, Dan Tufiş. *Resurse lingvistice pentru un sistem de întrebare-răspuns pentru limba română*, Revista Română de Interacţiune Om-Calculator 2 (2009), pp. 1–17.

[2] V. Barbu Mititelu. *Statistics on Derivation and its Representation in the Romanian Wordnet*, Proceedings of the 9th International Conference "Linguistic Resources and Tools for Processing the Romanian Language", pp. 99–108, 2013.

[3] O. Bilgin, O. Cetinoglu, K. Oflazer. *Morphosemantic relations in and across wordnets: A study based on Turkish*, Proceedings of GWC, pp. 60–66, 2004.

[4] C. Fellbaum (Ed.). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

[5] C. Fellbaum, P. Vossen. *The Challenge of Multilingual WordNets.* Lexical Resources and Evaluation 46, pp. 313–326, 2012.

[6] R. Ion, D. Tufiş. *Multilingual versus Monolingual Word Sense Disambiguation.* International Journal of Speech Technology, vol. 12, no 2-3 (2009), pp. 113–124.

[7] Radu Ion, Dan Ştefănescu, Alexandru Ceauşu, Dan Tufiş, Elena Irimia, Verginica Barbu Mititelu. *A Trainable Multi-factored QA System.* In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Pen as, Giovanna Roda (Eds.) Multilingual Information Access Evaluation, Vol. I Text Retrieval Experiments, pp. 257–264, Lecture Notes in Computer Science, Volume 6241/2010, Springer-Verlag.

[8] R. Ion, D. Ştefănescu. *Unsupervised Word Sense Disambiguation with Lexical Chains and Graph-Based Context Formalization.* LTC 2009 Proc. LNCS, vol. 6562 (2011), pp. 435–443.

[9] N. Kahusk, K. Kerner, K. Vider. *Enriching Estonian WordNet with Derivations and Semantic Relations.* Proceeding of the 2010 conference on Human Language Technologies – The Baltic Perspective, pp.195–200, 2010.

[10] S. Koeva. *Derivational and Morphosemantic Relations in Bulgarian Wordnet.* Intelligent Information Systems, XVI, Warsaw, Academic Publishing House, pp. 359–389, 2008.

[11] S. Koeva, C. Krstev, D. Vitas. *Morpho-semantic relations in Wordnet – a case study for two Slavic languages.* Proceedings of the Fourth Global WordNet Conference, Szeged, pp. 239–254, 2008.

[12] K. Linden, J. Niemi. *Is It Possible to Create a Very Large WordNet in 100 days? – an Evaluation*, Language Resources and Evaluation, 2013.

[13] G.A. Miller, R. Beckwith, C. Felbaum, D. Gross, K. Miller. *Five papers on WordNet*. Technical report, Cognitive Science Laboratory, Princeton University, August 1993. Revised version.

[14] G.A. Miller. *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, no. 11, pp. 39–41, 1995.

[15] K. Pala, D. Hlavackova. *Derivational relations in Czech Wordnet*. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, pp. 75–81, 2007.

[16] M. Piasecki, R. Ramocki, M. Maziarz. *Recognition of Polish Derivational Relations Based on Supervised Learning Scheme*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).

[17] H. Rodriguez, S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Alonge, F. Bertagna, A. Roventini. *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*. Computers and the Humanities, 32 (2-3), pp. 117–152, 1998.

[18] D. Tufiş, D. Cristea, S. Stamou. *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview*. Romanian Journal on Information Science and Technology. Special Issue on BalkaNet, volume 7, pp. 9–34, 2004.

[19] P. Vossen (Ed.). *EuroWordNet: A Multilingual Database with lexical Semantic Networks.* Kluwer. Dordrecht, The Netherlands.

Verginica Barbu Mititelu Received September 30, 2013

Research Institute for Artificial Intelligence "Mihai Drăgănescu",
Romanian Academy
13, Calea 13 Septembrie, Bucureşti 050711
Phone: +40-(0)213188103
E–mail: vergi@racai.ro

Computer Science Journal of Moldova, vol.**21**, no.3(63), 2013

Wiki-Translator: Multilingual Experiments for In-Domain Translations

Dan Tufiş, Radu Ion, Ştefan Daniel Dumitrescu

Abstract

The benefits of using comparable corpora for improving translation quality for statistical machine translators have been already shown by various researchers. The usual approach is starting with a baseline system, trained on out-of-domain parallel corpora, followed by its adaptation to the domain in which new translations are needed. The adaptation to a new domain, especially for a narrow one, is based on data extracted from comparable corpora from the new domain or from an as close as possible one. This article reports on a slightly different approach: building an SMT system entirely from comparable data for the domain of interest. Certainly, the approach is feasible if the comparable corpora are large enough to extract SMT useful data in sufficient quantities for a reliable training. The more comparable corpora, the better the results are. Wikipedia is definitely a very good candidate for such an experiment. We report on mass experiments showing significant improvements over a baseline system built from highly similar (almost parallel) text fragments extracted from Wikipedia. The improvements, statistically significant, are related to what we call the level of translational similarity between extracted pairs of sentences. The experiments were performed for three language pairs: Spanish-English, German-English and Romanian-English, based on sentence pairs extracted from the entire dumps of Wikipedia as of December 2012. Our experiments and comparison with similar work show that adding indiscriminately more data to a training corpus is not necessarily a good thing in SMT.

Keywords: comparable corpora, extraction of parallel sentences, language model, statistical machine translation, translation models.

1 Introduction

The research on domain adaptation based on comparable corpora has been motivated by the scarce parallel data for most of the language pairs or by the scarcity of (narrow) domain specific parallel data. The standard approach is to start with a baseline system, trained on as much as possible out-of-domain parallel corpora, followed by its adaptation to the domain in which new translations are needed. To date, OPUS[1] (Tiedemann, 2012) is the largest **online** collection of parallel corpora, comprising juridical texts (EUROPARL and EUconst), medical texts (EMEA), technical texts (e.g. software KDE manuals, PHP manuals), movie subtitles corpora (e.g. OpenSubs), translated transcribed talks (e.g. TED) or news (SETIMES) but these corpora are not available for all language pairs nor their sizes are similar with respect to the domain. Another example of a large collection of aligned parallel texts (in 22 languages) is JRC-Acquis (Steinberger et al., 2006), the total body of European Union (EU) law applicable in the EU Member States.

The adaptation to a new domain, especially for a narrow one, is based on data extracted from comparable corpora from the new domain or from an as close as possible one.

This article[2] reports on slightly different approach: building an SMT system entirely from comparable data for the domain of interest. The approach is feasible if the comparable corpora are large enough to extract SMT useful data in sufficient quantities for a reliable training. If the corpora, from which the translation-useful data are searched for, are strongly comparable, the outcomes may be surprisingly good.

Wikipedia is definitely a very good candidate for such an experiment. Wikipedia is not a real parallel corpus, but a strongly comparable multilingual corpus with many documents in different languages representing translations from (mainly) English. More often than not, the documents in one language are shortened or adapted translations of

[1] http://opus.lingfil.uu.se/

[2] A preliminary version of the results have been described in (Tufiş et al, 2013); here we bring more and new experimental results and comments.

documents from other (not always the same) languages and this property of Wikipedia together with its size makes it the ideal candidate of a strongly comparable corpus from which parallel sentences can be mined.

SMT engines like Moses[3] produce better translations when presented with larger and larger parallel corpora. In this context, large and good quality parallel corpora extracted from Wikipedia for different language pairs, can serve three purposes:

1. provide in-domain training data for aiding automatic translation of English (or other languages) Wikipedia articles into other languages thus paving the way to growing for poorer foreign Wikipedia sites;

2. provide in-domain training data for aiding automatic translation of Wikipedia non-English articles thus helping the dissemination of other nations' cultural and scientific contributions;

3. add a new domain (for many language pairs), the encyclopedic domain, to the list of domains for which parallel data already exist.

The structure of the article is as follows: we begin with a short review of related research (Section 2), continue with an informal description of the tool that was used to collect the parallel sentences from Wikipedia (Section 3). In Section 4, we describe the two-steps methodology for data extraction and provide quantitative data about the obtained parallel corpora for the English-Spanish, English-German and English-Romanian language pairs. We also provide BLEU evaluation of SMT using extracted data. Next, in Section 5 we compare our experiments with similar ones. The last section draws some conclusions and presents future plans.

[3]http://www.statmt.org/moses/

2 Related work

Adafre and Rijke (2006) were among the first to attempt extraction of parallel sentences from Wikipedia. Their approach consists of two experiments: 1) the use of an MT system (Babelfish) to translate from English to Dutch and then, by word overlapping, to measure the similarity between the translated sentences and the original sentences and 2) with an automatically induced (phrase) translation lexicon from the titles of the linked articles, they measure the similarity of source (English) and target (Dutch) sentences by mapping them to (multiple) entries in the lexicon and computing lexicon entry overlap. Experiments were performed on 30 randomly selected English-Dutch document pairs yielding a few hundred parallel sentence pairs.

Mohammadi and GhasemAghaee (2010) continue the work of Adafre and Rijke (2006) by imposing certain limits on the sentence pairs that can be formed from a Wikipedia document pair: the length of the parallel sentence candidates must correlate and the Jaccard similarity of the lexicon entries (seen as IDs) mapped to source (Persian) and target (English) must be as high as possible. As with Adafre and Rijke, the work performed by Mohammadi and GhasemAghaee does not actually generate a parallel corpus but only a couple hundred parallel sentences intended as a proof of concept.

Another experiment, due to Smith et al. (2010), addressed large-scale parallel sentence mining from Wikipedia. They automatically extracted large volumes of parallel English-Spanish (almost 2M pairs), English-German (almost 1.7M pairs) and English-Bulgarian (more than 145K pairs) sentences using binary Maximum Entropy classifiers (Munteanu and Marcu, 2005). The work of Smith et al. (2010) is the only one we are aware of, which extracted parallel corpora of similar sizes to ours. They released their Wikipedia test sets for English-Spanish (500 pairs) and for English-German (314 pairs), an inescapable opportunity for a direct comparison between our results and theirs. This comparison is documented in the Section 5.

3 Extracting bilingual comparable translation units

The EU project ACCURAT[4] (2010-2013) collected from the web very large sets of comparable documents and classified them (using a specially designed metrics) into different comparability classes: strongly comparable, comparable, weakly comparable and unrelated documents (see the public site for detailed reports and the associated data). From different comparability classes, various text mining systems, developed within the project, extracted useful MT data (highly similar cross-lingual sentences and parallel terms and name entities) which subsequently were used for assessing the impact on the translation quality of the existing baseline systems (Skadia et al., 2012; Tufiş, 2012).

For the experiments described in this article we used one of the ACCURAT text miners, namely LEXACC (Ştefănescu et al., 2012). It is a fast algorithm for parallel sentence mining from comparable corpora, developed to handle large amounts of comparable corpora in a reasonable amount of time. Unlike most text-miners based on binary classifiers (e.g. Munteanu and Marcu (2005)) which do not make the distinction between truly parallel sentences, partial parallel sentences, strongly comparable or weakly comparable sentences (or other, finer degrees of parallelism), LEXACC uses a similarity metrics allowing for ranking translation candidate pairs according to their similarity scores (with values continuously ranging from a very low number assigned to unrelated sentences to a very high number assigned to truly parallel sentences).

In order to significantly reduce the search space, LEXACC uses Lucene[5] to index the entire collection of target sentences (storing the document pair ID with each sentence). Using CLIR techniques the candidate sentence pairs are subject to several restricting filters (e.g. the length of the source and target sentence candidates must correlate, a high proportion of the source sentence content words must have a translation in the target candidate, etc.). In order to extract more in-

[4]http://www.accurat-project.eu/
[5]http://lucene.apache.org/

formative sentence pairs, LEXACC filters out titles or short sentences (with less than 3 words). Once this fast initial filtering is finished, a second step, computationally more expensive, generates the final similarity ranking of the translation pairs and leaves out all the pairs with a score below a pre-established threshold.

The translation similarity measure is a weighted sum of feature functions that indicate if the source piece of text is translated by the target. Given two sentences, s in the source language and t in the target language, then the translation similarity measure $P(s,t)$ is:

$$P(s,t) = \sum_i \theta_i f_i(s,t) \tag{1}$$

such that $\sum_i \theta_i = 1$. Each feature function $f_i(s,t)$ returns a real value between 0 (s and t are not related at all) and 1 (t is a translation of s) and contributes to the overall parallelism score with a specific fraction θ_i that is language-pair dependent and that is automatically determined by training a logistic regression classifier on existing parallel data in both directions: source-target and target-source. It follows that the translation similarity measure possible values are between 0 (s and t are not related at all) and 1 (t is a translation of s).

Some of the features used by the translation similarity measure in equation 1 are as follows (for a detailed description of these features, the reader is directed to (Ştefănescu et al., 2012)):

- the content words translation strength (i.e. the score of the best alignment between content words of s and t);
- the functional words translation strength (i.e. the score of the best alignment of functional words near a strong alignment link of content words);
- alignment obliqueness (i.e. the score of a content word alignment whose links do not cross is larger than the score of an alignment with crossing links);

- the *sentinel* translations feature: we noticed that, more often than not, two parallel sentences begin and end with strongly related content words even if words in the middle are not found in the lexicon.

In order to use LEXACC for mining useful MT sentence pairs one needs a translation lexicon. Ideally, this lexicon should be domain-specific, with a large lexical coverage of the search space. However, this requirement, difficult to meet, may be avoided using a boosting technique: use any available bilingual lexicon or extract a translation lexicon from whatever parallel corpora; run LEXACC on the in-domain comparable corpus and use the mined sentence pairs for extracting better in-domain lexicons; redo the sentence pair extraction. The boosting procedure may be repeated a number of times until no improvements are observed. Yet, one has to consider that the entire chain of processing is highly computational intensive, and depending on the size of the search space (as is the case of large Wikipedias) it may take several days.

4 Mining Wikipedia

Among the 285 language editions of Wikipedia (http://meta.wikimedia.org/wiki/List_of_Wikipedias[6]) created under the auspices of Wikimedia Foundation, the English, German and Spanish ones are listed in the best populated category, with more than 1,000,000 articles: English is the largest collection with 4,238,043 articles, German is the second largest with 1,587,660 articles while Spanish is the 6^{th} in the top Wikipedias with 1,017,938 articles. Romanian Wikipedia is in the medium populated category, and with 226,004 articles is the 25^{th} largest collection. For our experiments we selected three very large Wikipedias (English, German and Spanish) and a medium sized Wikipedia (Romanian) and performed SMT experiments for three language pairs: English-German, English-Spanish and English-Romanian.

[6]Consulted on May 22^{nd} 2013:

With these monolingual Wikipedias selected for parallel sentence mining, we downloaded (December 22^{nd}, 2012) the "database backup dumps"[7] for which Wikipedia states that they contain "a complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML". Parsing the English XML dump, we kept only the proper encyclopedic articles which contain links to their corresponding articles in the Spanish, German or Romanian. Thus, we removed articles that were *talks* (e.g. Talk:Atlas Shrugged), *logs* (e.g. Wikipedia:Deletion log), *user related articles* (e.g. User:AnonymousCoward), *membership related articles* (e.g. Wikipedia:Building Wikipedia membership), *manuals* and *rules related articles*, etc.

For each language, the retained articles were processed using regular expressions to remove the XML mark-up in order to keep only the raw, UTF-8 encoded text, which was saved into a separate file. The non-textual entries like images or tables were stripped off. Each text document was then sentence-split using an in-house freely available[8] sentence splitter based on a Maximum Entropy classifier.

Table 1. Linked documents for three language pairs

Language pair	**Document pairs**	**Size on disk**	**Size ratio (L1/L2)**
English-German	715,555	2.8 Gb (English)	1,22
		2.3 Gb (German)	
English-Romanian	122,532	778.1 Mb	3,91
		198.9 Mb	
English-Spanish	573,771	2.5 Gb	1,66
		1.5 Gb	

Table 1 lists the number of sentence-split Wikipedia comparable document pairs (by following the inter-lingual links) for each considered

[7] http://dumps.wikimedia.org/backup-index.html
[8] http://nlptools.racai.ro/nlptools/index.php?page=ssplit

language pair (see (Ştefănescu and Ion, 2013) for further details).

Looking at the size ratio of the linked documents for each language pair it is apparent that Romanian documents are much shorter than the linked English ones. The size ratios for other language pairs are more balanced, coming closer to expected language specific ratio for a parallel corpus (see below).

We applied the boosting procedure as follows:

a) We used the JRC-Acquis parallel corpora to extract initial translation lexicons for English-Romanian and English-German language pairs. For English-Spanish pair we used the corresponding parallel sub-part of EUROPARL. We run GIZA++ (Gao and Vogel, 2008) and symmetrized the extracted translation lexicons between the source and target languages. The Romanian-English lexicon extracted with GIZA++ was merged with an in-house dictionary generated from our wordnet (Tufiş et al., 2013) aligned to Princeton WordNet. With these lexicons we performed the first phase of LEXACC extraction of comparable sentence pairs from the respective Wikipedias. Let us call this data-set, for a language pair L1-L2, as Wiki-Base (L1, L2). The experiments with Wiki-Base for three language pairs are described in Section 4.2;

b) From the most MT useful parts of Wiki-Base(L1, L2), as resulted from the first step, we extracted new translation lexicons used for a second phase (the boosting) of LEXACC (symmetrized) extraction, thus getting a new and larger data set which we refer to as Wiki-Train (L1,L2). The most useful parts of Wiki-Train were identified based on their impact on the BLEU score for the test set as described in Section 4.3 and used for the training of the Wiki-Translators.

4.1 Building Wiki-Base (L1,L2)

Table 2 lists, for different similarity scores as extraction thresholds, the number of MT useful sentence pairs (P) found in each language pair dataset, as well as the number of words (ignoring punctuation) per

language (English Words, German Words, Romanian Words, Spanish Words) in the respective sets of sentence pairs. Obviously, data extracted with a given Similarity score threshold was a proper sub-set of any data extracted with a lower Similarity score threshold.

Table 2: Wiki-base: number of parallel sentences and words for each language pair, for a given threshold

Similarity score	**English-Romanian**	**English-German**	**English-Spanish**
0.9	Pairs: 42,201 English Words: 0.814 M Romanian Words: 0.828 M	Pairs: 38,390 English Words: 0.554 M German Words: 0.543 M	Pairs: 91,630 English Words: 1.126 M Spanish Words: 1.158 M
0.8	Pairs: 112,341 English Words: 2.356 M Romanian Words: 2.399 M	Pairs: 119,480 English Words: 2.077 M German Words: 2.010 M	Pairs: 576,179 English Words: 10.504 M Spanish Words: 11.285 M
0.7	Pairs: 142,512 English Words: 2.987 M Romanian Words: 3.036 M	Pairs: 190,135 English Words: 3.494 M German Words: 3.371 M	Pairs: 1,219,866 English Words: 23.730 M Spanish Words: 25.931 M
0.6	Pairs: 169,662 English Words: 3.577 M Romanian Words: 3.634 M	Pairs: 255,128 English Words: 4.891 M German Words: 4.698 M	Pairs: 1,579,692 English Words: 31.022 M Spanish Words: 33.706 M

Continuation of Table 2

Similarity score	**English-Romanian**	**English-German**	**English-Spanish**
0.5	Pairs: 201,263 English Words: 4.262 M Romanian Words: 4.325 M	Pairs: 322,011 English Words: 6.453 M German Words: 6.186 M	Pairs: 1,838,794 English Words: 36.512 M Spanish Words: 39.545 M
0.4	Pairs: 252,203 English Words: 5.415 M Romanian Words: 5.482 M	Pairs: 412,608 English Words: 8.470 M German Words:8.132 M	Pairs: 2,102,025 English Words: 42.316 M Spanish Words: 45.565 M
0.3	Pairs: 317,238 English Words: 6.886 M Romanian Words: 6.963 M	Pairs: 559,235 English Words: 11.797 M German Words: 11.353M	Pairs: 2,656,915 English Words: 54.932 M Spanish Words: 58.524 M

Depending on the similarity threshold, the extracted pairs of sentences may be really parallel, may contain real parallel fragments, may be similar in meaning but with a different wording, or lexically unrelated in spite of domain similarity. That is, the lower the threshold, the higher the noise.

By random manual inspection of the generated sentence pairs, we saw that, in general, irrespective of the language pair, sentence pairs with a translation similarity measure higher than 0.6 are parallel. Based on the number of words in each language side of the parallel extracted sentences, one can easily compute an expected average length ratio for the three considered language pairs. Those pairs with a translation similarity measure of at least 0.5 have extended parallel fragments which an accurate word or phrase aligner easily detects.

Further down the threshold scale, below 0.3, we usually find sentences that roughly speak of the same event but are not actual translations of each other. The noisiest data sets were extracted for the 0.1 and 0.2 similarity thresholds and we drop them from further experiments.

If we consider the extraction rate ExtR as the ratio between the number of parallel sentences (those with similarity score higher or equal to 0.7) and the number of linked documents we get the following figures:

$$\mathrm{ExtR}(\mathrm{En}-\mathrm{Ro}) = 1.16; \quad \mathrm{ExtR}(\mathrm{En}-\mathrm{De}) = 0.26;$$

$$\mathrm{ExtR}(\mathrm{En}-\mathrm{Es}) = 2.12.$$

The striking differences may have several explanations. The first one is that the Spanish documents linked to English documents are more literary translated, while the German documents are more distant from the English documents to which they are linked. Romanian documents are somewhere in between. Another partial explanation might be the quality of the dictionaries LEXACC used for each language. Augmenting the Romanian-English lexicon extracted by GIZA++ from JRC-Acquis with the data from wordnet resulted in a cleaner (although smaller) dictionary than the German-English extracted also from JRC-Acquis. In case of Spanish-English extraction rate (higher than for other two language pairs) we hypothesize that the GIZA++ dictionary extracted from EUROPARL has a better covering of the Wikipedia vocabulary. The experimental results described in the following sections strongly support these hypotheses.

4.2 SMT experiments with Wiki-Base

In order to select the most MT useful parts of Wiki-Base for the three considered language pairs, we built three baseline Moses-based SMT systems using only parallel sentences, that is those pairs extracted with a similarity score higher or equal to 0.7 (see Table 2). We incrementally extended the training data by lowering the similarity score threshold and, using the same test-set, observed the variation of the BLEU score. The purpose for the evaluation of the SMT systems was only to indicate what would be the best threshold for selecting the

training set from the Wiki-Train for building the Wiki-Translators. As the standard SMT system we chose Moses surface-to surface translation and lexical reordering model with parameters wbe-msd-bidirectional-fe, with phrase-length of maximum 4 words, and the default values for the rest of parameters.

The **language model** (LM) for all experiments was trained on all monolingual, sentence-split English Wikipedia after removing the administrative articles as described in Section 3. The language model was limited to 5-grams and the counts were smoothed by the interpolated Knesser-Ney method.

Since we experimentally noticed that the additional sentence pairs extracted for a threshold of 0.6 were almost as parallel as those extracted for higher thresholds we included this interval too in the sampling process for **test set** design. Thus, we proceeded to randomly sample 2,500 sentence pairs from similarity intervals ensuring parallelism ([0.6, 0.7), [0.7-0.8), [0.8, 0.9) and [0.9-1]). We obtained 10,000 parallel sentence pairs for each language pair. Additionally, we extracted 1,000 sentence pairs as development set (**dev set**). These 11,000 sentences were removed from the training corpora of each language pair. When sampling parallel sentence pairs, we were careful to obey the Moses' filtering constraints: both the source and target sentences must have at least 4 words and at most 60 words and the ratio of the longer sentence (in tokens) of the pair over the shorter one must not exceed 2. The duplicates were also removed.

Further on, we trained **seven translation models** (TM), for each language pair, over cumulative threshold intervals beginning with 0.3: TM_1 for [0.3, 1], TM_2 for [0.4, 1] ..., TM_7 for [0.9, 1]. The resulting eight training corpora have been filtered with Moses' cleaning script with the same restrictions mentioned above. For every language, both the training corpora and the test set have been tokenized using Moses' tokenizer script and true-cased. The quality of the translation systems is measured as usual in terms of their BLUE score (Papineni et al., 2002) on the same test data.

We have to emphasize that the removal of the sentences in the test and development sets from the training corpora does not ensure an un-

biased evaluation of the BLEU scores since their context still remained in the training corpora. This requires some more explanations. For each extracted sentence pair, LEXACC stores in a book-keeping file, the ID of the document-pair out of which the extraction was done. This information allows for elimination from the training set of all the pairs coming from the same documents from which the development and evaluation sets were selected. However, due to the nature of the Wikipedia article authoring, this strategy of filtering the development and evaluation sets does not ensure an unbiased evaluation. The Wikipedia contributors are given specific instructions for authoring documents[9] and by observing these instructions, inherently one could find in different documents almost identical sentences except for a few name entities. Indeed we found examples of such sentence pairs in the train set similar, but not identical, to sentences in the test set, yet coming from different document-pairs. Certainly one could build a tough test-set by removing from train set all similar (pattern-based) sentences, but we did not do that because it would have been beyond the purpose of this work. As we mentioned before, this evaluation was meant only for estimating most useful extraction level for the second phase of training the WIKI-Translators.

Table 3 summarizes the results of this first step experiment, the bold characters identifying the most MT useful parts of Wiki-Base (L1,L2). We considered TM $_{[0.7,1]}$ (the shaded line in Table 3) as the baseline for all language pairs.

4.3 Building Wiki-Train (L1,L2)

The experiments on Wiki-base revealed that the most useful training data has been extracted by using LEXACC with 0.5 similarity score for German-English and Romanian-English language pairs and 0.3 for Spanish-English pair (see Table 3). We re-run GIZA++ on these subsets of Wiki-Base to extract new in-domain lexicons.

The new lexicons were merged with the initial ones and the LEXACC extraction was repeated with the resulted mined comparable

[9] http://en.wikipedia.org/wiki/Wikipedia:Translation

Table 3. Comparison between SMT systems trained on various parts of Wiki-Base

TM based on Wiki-Base	**BLEU SCORE Romanian –> English**	**BLEU SCORE German –> English**	**BLEU SCORE Spanish –> English**
$TM_{[0.3,1]}$	37.24	39.16	**47.59**
$TM_{[0.4,1]}$	37.71	39.46	47.52
$TM_{[0.5,1]}$	**37.99**	**39.52**	47.53
$TM_{[0.6,1]}$	37.85	39.5	47.44
$TM_{[0.7,1]}$	37.39	39.24	47.28
$TM_{[0.8,1]}$	36.89	38.57	46.27
$TM_{[0.9,1]}$	32.76	34.73	39.68

sentence-pairs denoted as Wiki-Train.

As the experiments on the Wiki-base showed that for a similarity threshold less than or equal to 0.2 LEXACC delivers not very useful data, we started the second step of mining using the similarity scores of at least 0.3.

Table 4 shows the results of the boosted extraction process. As one can see the extracted data, at each similarity score level, is significantly increased for the English-Romanian and English-German language pairs. For English-Spanish, except for the similarity scores 0.8 and 0.9 the number of sentence pairs is smaller than in Wiki-Base. The reason is that in this round we detected several identical pairs with those in the training and development sets and several duplicated pairs in the training set. Anyway, the English-Spanish Wiki-Train was the largest train-set and containing the highest percentage of fully parallel sentence pairs.

Table 4: Wiki-Train: number of parallel sentences and words for each language pair, for a given threshold

Similarity score	**English-Romanian**	**English-German**	**English-Spanish**
0.9	Pairs: 66,777 English Words: 1.077 M Romanian Words: 1.085 M	Pairs: 97,930 English Words: 1.069 M German Words: 1.042 M	Pairs: 113,946 English Words: 1.164 M Spanish Words: 1.193 M
0.8	Pairs: 152,015 English Words: 2.688 M Romanian Words: 2.698 M	Pairs: 272,358 English Words: 3.695M German Words: 3.552 M	Pairs: 597,992 English Words: 9.733 M Spanish Words: 10.510 M
0.7	Pairs: 189,875 English Words: 3.364 M Romanian Words: 3.372 M	Pairs: 434,019 English Words: 6.201 M German Words: 5,929 M	Pairs: 1,122,379 English Words: 19.941 M Spanish Words: 21.821 M
0.6	Pairs: 221,661 English Words: 3.961 M Romanian Words: 3.970 M	Pairs: 611,868 English Words: 8.944 M German Words8.532 M	Pairs: 1,393,444 English Words: 25.068 M Spanish Words: 27.411 M
0.5	Pairs: 260,287 English Words: 4,715 M Romanian Words: 4,722 M	Pairs: 814,041 English Words: 12.361 M German Words: 11.792M	Pairs: 1,587,276 English Words: 28.987 M Spanish Words: 31.567 M

Continuation of Table 4

Similarity score	**English-Romanian**	**English-German**	**English-Spanish**
0.4	Pairs: 335,615 English Words: 6.329 M Romanian Words: 6.324 M	Pairs: 1,136,734 English Words: 18,089 M German Words: 17.306 M	Pairs: 1,807,892 English Words: 33.619 M Spanish Words: 36,369 M
0.3	Pairs: 444,102 English Words: 8.712 M Romanian Words: 8.700 M	Pairs: 1,848,651 English Words: 31.405 M German Words: 30.175 M	Pairs: 2,288,163 English Words: 44.021 M Spanish Words: 47.180 M

4.4 SMT experiments with Wiki-Train

The Wiki-Train corpora were used with the same experimental setup as described in Section 4.2. The training of each translation system was followed by the evaluation on the respective test sets (10,000 pairs) in both translation directions. To make the comparison between the translation qualities we did the translations without MERT optimization of the parameters. The results are presented in Table 5.

Having much more training data, in case of the Romanian –> English and German –>English the BLEU scores significantly increased (with 3.1 and 2.58 points respectively). For Spanish-English the decrease of number of sentences in Wiki-Train as compared to Wiki-Base negatively impacted the new BLEU score, which is 1.31 points lower. It would be interesting to see what would happen with a higher threshold training set, for instance $TM_{[0.5,1]}$, as used for the other language pairs.

As expected, the translations into non-English languages are less accurate due to a more complex morphology of the target language (most of the errors are morphological ones), but still the BLEU scores

are very high, better than most of the results we are aware off (for in-domain experiments).

Table 5. Best translation SMT systems, trained on Wiki-Train[10]

TM based on Wiki-Train	**$TM_{[0.5,1]}$ Romanian –> English**	**$TM_{[0.5,1]}$ German –> English**	**$TM_{[0.3,1]}$ Spanish –> English**
BLEU SCORE	**41.09**	**40.82**	**46.28**
	$TM_{[0.5,1]}$ English –> Romanian	**$TM_{[0.5,1]}$ English –> German**	**$TM_{[0.3,1]}$ English –> Spanish**
BLEU SCORE	**29.61**	**35.18**	**46.00**

5 Comparison with other works

Translation for Romanian-English language pair has also been studied in (Boroş et al., 2013; Dumitrescu et al., 2012; 2013) among others. In these works we had explicit interests in experiments on using in-domain/out-of-domain test/train data, and various configurations of the Moses decoder in surface-to-surface and factored translation. Out of the seven domain-specific corpora (Boroş et al., 2013) one was based on Wikipedia. The translation experiments on English-Romanian, similar to those reported here, were surface based (t0-0, m0) with training on parallel sentence pairs extracted from Wikipedia by LEXACC at a fixed threshold: 0.5 (called "WIKI5"), without MERT optimization. A random selection of unseen 1,000 Wikipedia Romanian test sentences[11] has been translated into English using combinations of:

- a WIKI5-based translation model (240K sentence pairs)/WIKI5-based language model;

[10]For a fair comparison with data in Table 3 we did not use here the MERT optimization

[11]The test-set construction followed the same methodology described in this article

- a global translation model (1.7M sentence pairs)/global language model named "ALL", made by concatenating all specific corpora.

Table 6 gives the BLEU scores for the Moses configuration similar to ours.

Table 6. BLEU scores on 1000 sentences Wikipedia test set of Dumitrescu et al. (2013)

	WIKI5 TM	**ALL TM**
WIKI5 LM	**29.99**	29.95
ALL LM	29.51	29.95

Boroş et al.'s results confirm the conclusion we claimed earlier: the ALL system performs worse than the in-domain WIKI5 system. The large difference between the herein BLEU score (41.09) and 29.99 in (Boroş et al., 2013) may be explained by various factors. First and more importantly, our current language model was entirely in-domain for the test data and much larger: the language model was built from entire Romanian Wikipedia (more than 220,000 documents) while the language model in (Boroş et al., 2013) was built only from the Romanian sentences paired to English sentences (less than 240,000 sentences). Our translation model was built from more than 260,000 sentence pairs versus 234,879 sentence pairs of WIKI5). Another explanation might be the use of different Moses filtering parameters (e.g. the length filtering parameters) and different test sets. As suggested by other researchers, Wikipedia-like documents are more difficult to translate than, for instance, legal texts. The BLEU scores on JRC-Acquis test sets (with domain specific training) reported in (Boroş et al., 2013) is almost double than those obtained on Wikipedia test sets.

The most similar experiments to ours have been reported by Smith et al. (2010). They mined for parallel sentences from Wikipedia producing parallel corpora of sizes even larger than ours. While they used for training all the extracted sentence pairs, we used only those subsets that observed a minimal similarity score. We checked to see if their test sets for English-Spanish (500 pairs) and for English-German

(314 pairs) contained sentences in our training sets and, as this was the case, we eliminated from the training several sentence pairs (about 200 sentence pairs from the English-Spanish training corpus and about 140 sentence pairs from the English-German training corpus). We re-trained the two systems on the slightly modified training corpora. Since in their experiments they used MERT-optimized translation systems, we optimized, also by MERT, our new $TM_{[0.5,1]}$ for German–>English and new $TM_{[0.3,1]}$ for Spanish–>English translation systems, using the respective dev-sets (each containing 1,000 sentence pairs).

Their test sets for English-Spanish and for English-German were translated (after being true-cased) with our best translation models and also with Google Translate (as of mid-February 2013).

Table 7 summarizes the results. In this table, "Large+Wiki" denotes the best translation model of Smith et al. which was trained on many corpora (including Europarl and JRC Acquis) and on more than 1.5M parallel sentences mined from Wikipedia. "$TM_{[0.4,1]}$" and "$TM_{[0.5,1]}$" are our Wiki-Train translation models as already explained. "Train data size" gives the size of training corpora in multiples of 1,000 sentence pairs.

Table 7. Comparison between SMT systems on the Wikipedia test set provided by Smith et al. (2010)

Language pair	**Train data size (sentence pairs)**	**System**	**BLEU**
Spanish-English	9,642K	Large+Wiki	43.30
	2,288K	$TM_{[0.4,1]}$	**50.19**
	–	Google	44.43
German-English	8,388K	Large+Wiki	23.30
	814K	$TM_{[0.5,1]}$	24.64
	–	Google	21.64

For Spanish-English test set of Smith et al. (2010) our result is significantly better than theirs, in spite of almost 4 times less training

data. For the German-English pair, the difference is larger between $TM_{[0.5,1]}$ and Large+Wiki systems, and one should also notice that our system used 10 times less training data (but, presumably, much cleaner).

However, our $TM_{[0.5,1]}$ for German-English performed on the new test set much worse than on our test-set (24.64 versus 40.82[12] BLEU points) which was not the case for the Spanish-English language pair. We suspected that some German-English translation pairs in the Smith et al. (2010) test set were not entirely parallel. This idea was supported by the correlation of the evaluation results between our translations and Google's for Spanish-English and German-English. Also, their reported results on German-English were almost half of the ones they obtained for Spanish-English.

Therefore, we checked the German-English and Spanish-English test sets (supposed to be parallel) by running the LEXACC miner to see the similarity scores for the paired sentences. The results confirmed our guess. The first observation was that the test sets contained pieces of texts that looked like section titles (e.g. BT: Contaminación – BT: Pollution; Segunda clase – Second class; Autoengano-Self-deception, in Spanish – English test-set or Städte – Cities and towns; 1956 Armagnac – 1956 Armagnac; Produkte – Products; Geschwindigkeitsrekorde – Speed records; Geschichte – History in the German-English test-set). Such short sentences were ignored by LEXACC. While out of the considered sentence pairs (ignoring the sentences with less than 3 words), for Spanish-English LEXACC identified more than 92% as potentially useful SMT pairs (with a similarity score higher than or equal to 0.3 – this was the extraction threshold for Spanish-English sentence-pairs), for German-English LEXACC identified only 35% potentially useful SMT pairs (a similarity score higher than or equal to 0.5 – this was the extraction threshold for German-English sentence-pairs). Even if the threshold for German-English was lowered to 0.3, only 45% passed the LEXACC filtering. As for parallelism status of the sentence pairs in the

[12]Note that this value for our $TM_{[0.5,1]}$ was obtained on a very different and much larger test set and also without MERT optimization. Yet, the difference is large enough to raise suspicions on the test set used for this comparison.

test-sets (i.e. similarity scores higher than 0.6 for both languages) the percentages were 78% for Spanish-English and only 29% for German-English. Without ignoring the short sentences (easy to translate) these percentages would have probably been a little bit higher (80.8% for Spanish-English and 32.82% for German-English).

These evaluations outline also that LEXACC is too conservative in its rankings: we noticed almost parallel sentences in the test-set for Spanish-English even for a similarity score of 0.1 while in the German-English the same happens for similarity scores lower than 0.3. The most plausible explanation was that one of the LEXACC's parameters (cross-linking factor) strongly discourages long-distance reordering (which was quite frequent in the German-English test set and has also a few instances in the Spanish-English test set).

Table 8 shows some examples of sentence pairs in the German-English and Spanish-English test sets showing low level of parallelism (inappropriate for translation quality evaluation) but also some examples of sentence pairs which were conservatively lower ranked by LEXACC.

Table 8: Examples of sentence-pairs in the German-English and Spanish-English test sets used by Smith et al. (2010)

Similarity	**German source sentence**	**English reference translation**
1	**2**	**3**
< 0.1	Zuletzt stand sie für Robert Dornhelms Historienfilm Kronprinz Rudolf als Mary von Vetsera und in Le Ragazze di San Frediano als Mafalda vor der Filmkamera.	Puccini 's role as Mafalda in the 2007 Rai Uno miniseries Le ragazze di San Frediano cast her among many other well-known Italian actresses, including Martina Stella, Chiara Conti, and Camilla Filippi.

Continuation of Table 8

1	**2**	**3**
< 0.1	Unter anderem ist es für die Durchführung der Volkszählung zuständig.	Every 10 years, this organisation conducts a national census.
< 0.1	Daraufhin nahm sich Nikolaus, der es mit der ehelichen Treue schon mehrfach nicht so genau genommen hatte, eine Mätresse, Alexandras Hofdame Barbara Nelidowa.	Nelidova went with them, and though Alexandra was jealous in the beginning, she soon came to accept the affair, and remained on good terms with her husband's mistress.
0.12	Im Unterschied zu Cognac wird Armagnac in einem kontinuierlichen Brennverfahren nur einmal destilliert, also nicht rektifiziert.	Armagnac is traditionally distilled once, which results initially in a less polished spirit than Cognac, where double distillation usually takes place.
0.29	Die 64,5 Prozent , welche die SPD unter seiner Führung erzielte, waren das höchste Ergebnis, welches je eine Partei auf Bundeslandsebene bei einer freien Wahl in Deutschland erzielt hatte.	In the election that was conducted in the western part of Berlin two months later, his popularity gave the SPD the highest win with 64.5 % ever achieved by any party in a free election in Germany.
Similarity	**Spanish source sentence**	**English reference translation**
Misaligned	En febrero de 1988, a 12 UA del Sol, el brillo de Quirón alcanzó el 75 % Este comportamiento es típico de los cometas pero no de los asteroides.	In February 1988, 12 AU from the Sun, Chiron brightness reached 75%.

Continuation of Table 8

1	2	3
0.1	Sin embargo, el museo, llamado no fue terminado sino hasta el 10 de abril de 1981, dos días antes del vigésimo aniversario del vuelo de Yuri Gagarin.	However, it took until April 10, 1981 (two days before the 20th anniversary of Yuri Gagarin's flight) to complete the preparatory work and open the Memorial Museum of Cosmonautics.

6 Conclusions

Wikipedia is a rich resource for parallel sentence mining in SMT. Comparing different translation models containing MT useful data ranging from strongly comparable, to parallel, we concluded that there is sufficient empirical evidence not to dismiss sentence pairs that are not fully parallel on the suspicion that because of the inherent noise they might be detrimental to the translation quality. On the contrary, our experiments demonstrated that in-domain comparable data are strongly preferable to out-of-domain parallel data. However, there is an optimum level of similarity between the comparable sentences, which according to our similarity metrics (for the language pairs we worked with) is around 0.4 or 0.5.

Additionally, the two step procedure we presented, demonstrated that an initial in-domain translation dictionary is not necessary, it can be constructed subsequently, starting with a dictionary extracted from whatever out-of-domain data.

We want to mention that it is not the case that our extracted Wikipedia data is the maximally MT useful data. First of all, LEXACC may be improved in many ways, which is a matter for future developments. For instance, although the cross-linking feature is highly relevant for language pairs with similar word ordering, it is not very effective for language pairs showing long distance re-ordering. We also noticed that a candidate pair for which its two parts contained different numerical entities (numbers, dates, times) was dropped from further

consideration. Thirdly, the extraction parameters of LEXACC were not re-estimated for the Wiki-Train construction. Additionally, we have to mention that LEXACC evaluated and extracted only full sentences: a finer-grained (sub-sentential) extractor would likely generate more MT useful data. Also, one should note that the evaluation figures are just indicative for the potential of Wikipedia as a source for SMT training. In previous work it was shown that using factored models for inflectional target languages (Boroş et al, 2013) and cascading translators (Tufiş and Dumitrescu, 2012) may significantly improve (several BLEU points) the translation accuracy of an SMT system. Some other techniques may be used to improve at least translations into English. For instance, given that English adjectives and all functional words are not inflected, a very effective way, for a source inflectional language would be to lemmatize all words in these categories. Another idea is to split compound words of a source language (such as German) into their constituents. Both such simplifications are, computationally, not very expensive (and for many languages appropriate tools are publicly available) but may significantly reduce the number of out-of-vocabulary input tokens.

The parallel Wiki corpora (before and after the boosting step), including the test sets (containing 10,000) and the dev–sets (containing 1,000 sentences) are freely available on-line[13].

Acknowledgments. This work has been supported by the EU under the Grant Agreements no. 248347 (ACCURAT) and no. 270893 (METANET4U).

References

[1] Sisay Fissaha Adafre and Maarten de Rijke. 2006. *Finding similar sentences across multiple languages in Wikipedia.* In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), April 3-7, 2006. Trento, Italy, pp. 62–69.

[13] http://ws.racai.ro:9191/repository/search/

[2] Tiberiu Boroş, Stefan Daniel Dumitrescu, Radu Ion, Dan Ştefănescu, Dan Tufiş. 2013. *Romanian-English Statistical Translation at RACAI.* În E. Mitocariu, M. A. Moruz, D. Cristea, D. Tufiş, M. Clim (eds.) Proceedings of the 9th International Conference "Linguistic Resources and Tools for Processing the Romanian Language", 16-17 mai, 2013, Miclăuşeni, Romania, 2013. "Alexandru Ioan Cuza" University Publishing House. 197 p. ISSN 1843-911X. pp. 81–98, 2013.

[3] Ştefan Dumitrescu, Radu Ion, Dan Ştefănescu, Tiberiu Boroş, Dan Tufiş. 2013. *Experiments on Language and Translation Models Adaptation for Statistical Machine Translation.* In Dan Tufiş, Vasile Rus, Corina Forăscu (eds.) *Towards Multilingual Europe 2020: A Romanian Perspective,* pp. 205–224, 2013.

[4] Ştefan Dumitrescu, Radu Ion, Dan Ştefănescu, Tiberiu Boroş, Dan Tufiş. *Romanian to English Automatic MT Experiments at IWSLT12.* In Proceedings of the International Workshop on Spoken Language Translation, December 6 and 7, 2012, Hong Kong, pp. 136–143.

[5] Qin Gao and Stephan Vogel. 2008. *Parallel implementations of a word alignment tool.* In Proceedings of ACL-08 HLT: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, June 20, 2008. The Ohio State University, Columbus, Ohio, USA, pp. 49–57.

[6] Philipp Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation,* In Proceedings of the tenth Machine Translation Summit, Phuket, Thailand, pp. 79–86.

[7] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. *Moses: open source toolkit for statistical machine translation.* In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and

Demonstration Sessions (ACL '07). Association for Computational Linguistics, Stroudsburg, PA, USA, pp.177–180.

[8] Mehdi Mohammadi and Nasser GhasemAghaee. 2010. *Building bilingual parallel corpora based on Wikipedia.* In Computer Engineering and Applications (ICCEA 2010), Second International Conference on Computer Engineering and Applications, Vol. 2, pp. 264–268. IEEE Computer Society Washington, DC, USA.

[9] Dragoş Munteanu, Daniel Marcu. 2005. *Improving machine translation performance by exploiting comparable corpora.* Computational Linguistics, 31(4), pp. 477–504.

[10] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. *BLEU: A method for automatic evaluation of machine translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), July 2002. Philadelphia, USA, pp. 311–318.

[11] Inguna Skadiņa, Ahmet Aker, Nikos Glaros, Fangzhong Su, Dan Tufiş, Mateja Verlic, Andrejs Vasiljevs, Bogdan Babych, 2012. *Collecting and Using Comparable Corpora for Statistical Machine Translation.* In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), May 23-26, 2012. Istanbul, Turkey, ISBN 978-2-9517408-7-7.

[12] Jason R. Smith, Chris Quirk, Kristina Toutanova. 2010. *Extracting parallel sentences from comparable corpora using document level alignment.* In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 403–411. © Association for Computational Linguistics (2010).

[13] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş. 2006. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.* In Proceedings of the 5^{th} LREC Conference, Genoa, Italy, 22-28 May, 2006, ISBN 2-9517408-2-4, EAN 978-2-9517408-2-2.

[14] Dan Ştefănescu, Radu Ion, Sabine Hunsicker. 2012. *Hybrid parallel sentence mining from comparable corpora.* In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), pp. 137–144, Trento, Italy, May 28-30, 2012.

[15] Jörg Tiedemann. 2012. *Parallel data, tools and interfaces in OPUS.* In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), May 23-26, 2012. Istanbul, Turkey, ISBN 978-2-9517408-7-7.

[16] Dan Tufiş. *Finding Translation Examples for Under-Resourced Language Pairs or for Narrow Domains; the Case for Machine Translation.* 2012. In Computer Science Journal of Moldova, Academy of Sciences of Moldova, Institute of Mathematics and Computer Science, ISSN 1561-4042, vol.20, no.2(59), pp. 227–245.

[17] Dan Tufiş, Verginica Barbu Mititelu, Dan Ştefănescu, Radu Ion. 2013. *The Romanian Wordnet in a Nutshell.* Language and Evaluation, Springer, Vol. 47, no. 2, 2013, ISSN 1574-020X, DOI: 10.1007/s10579-013-9230-7

[18] Dan Tufiş, Radu Ion, Ştefan Dumitrescu, Dan Ştefănescu. 2013. *Wikipedia as an SMT Training Corpus.* In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, September 7-13, 2013.

[19] Dan Ştefănescu, Radu Ion. 2013. *Parallel-Wiki: A Collection of Parallel Sentences Extracted from Wikipedia.* In Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics, pp. 137–144, Trento, Italy, March 24-30, 2013, Samos, Greece.

Dan Tufiş[1], Radu Ion[2], Ştefan Daniel Dumitrescu[3] Received September 16, 2013

Institute for AI, Romanian Academy, Bucharest, Romania
[1] E-mail: tufis@racai.ro
[2] E-mail: radu@racai.ro
[3] E-mail: danstef@racai.ro

Computer Science Journal of Moldova, vol.**21**, no.3(63), 2013

Intelligent Robust Robotic Controllers: SW & HW Toolkit of Applied Quantum Soft Computing

Sergey V. Ulyanov, Andrey G. Reshetnikov, Timur A. Kerimov

Abstract

A generalized design strategy of intelligent robust control systems based on quantum/soft computing technologies that enhance robustness of hybrid intelligent controllers by supplying a self-organizing capability is described. We stress our attention on the robustness features of intelligent control systems in unpredicted control situations with the simulation of Benchmark.

Keywords: Intelligent Control, Quantum algorithm, Self-Organization, Knowledge Base, Quantum Fuzzy Inference.

1 Introduction

For complex and ill-defined dynamic control objects that are not easily controlled by conventional control systems (such as P-[I]-D-controllers) – especially in the presence of fuzzy model parameters and different stochastic noises – the System of Systems Engineering methodology provides fuzzy controllers (FC) as one of alternative way of control systems design.

Soft computing methodologies, such as genetic algorithms (GA) and fuzzy neural networks (FNN) had expanded application areas of FC by adding optimization, learning and adaptation features.

But still now it is difficult to design optimal and robust intelligent control system, when its operational conditions have to evolve dramatically (aging, sensor failure and so on).

Such conditions could be predicted from one hand, but it is difficult to cover such situations by a single FC.

Using unconventional computational intelligence toolkit we propose a solution of such kind of generalization problems by introducing a *self-organization* design process of robust KB – FC that is supported by the *Quantum Fuzzy Inference* (QFI) based on quantum soft computing ideas [1–3].

2 Problem's Formulation

2.1 Main problem and toolkit

One of the main problems in modern FC design is how to design and introduce robust KBs into control system for increasing *self-learning, self-adaptation and self-organizing capabilities* that enhance robustness of developed FC in unpredicted control situations.

The *learning* and *adaptation* aspects of FC's have always the interesting topic in advanced control theory and system of systems engineering. Many learning schemes were based on the *back-propagation* (BP) algorithm and its modifications (see, for example, [1] and their references). Adaptation processes are based on iterative stochastic algorithms.

These ideas are successfully working if we perform our control task without a presence of ill-defined stochastic noises in environment or without a presence of unknown noises in sensors systems and control loop, and so on.

For more complicated control situations learning and adaptation methods based on BP-algorithms or iterative stochastic algorithms do not guarantee the required robustness and accuracy of control.

The solution of this problem based on Soft Computing Optimizer (SCO) was developed in [2]. For achieving of *self-organization* level in intelligent control system it is necessary to use QFI [3, 4]. The described *self-organizing* FC design method is based on special form of QFI that uses a few of partial KBs designed by SCO.

QFI uses the laws of quantum computing technologies [5] and ex-

plores three main unitary operations: (i) superposition; (ii) entanglement (quantum correlations); and (iii) interference. According to quantum gate computation, the logical union of a few KBs in one generalized space is realized with *superposition* operator; with *entanglement* operator (that can be equivalently described by different models of *quantum oracle* [6]) a search of a "successful" marked solution is formalized; and with *interference* operator we can extract "good" solutions with classical *measurement* operations [7].

2.2 Method of solution

The proposed QFI system consists of a few KB-FCs, each of which has prepared for appropriate conditions of control object and excitations by SCO [2]. QFI system is a new quantum control algorithm of self-organization block, which performs post processing of the results of fuzzy inference of each independent FC and produces in on-line the generalized control signal output [4].

In this case the output of QFI is an optimal robust control signal, which combines best features of each independent FC outputs. Therefore the operation area of such a control system can be expanded greatly as well as its robustness.

Robustness of control is the background for support the reliability of advanced control accuracy in uncertainty and information risk [5].

The simulation example of robust intelligent control based on QFI is introduced.

2.3 Main goal

The main technical purpose of QFI is to supply a self-organization capability for many (sometimes unpredicted) control situations based on a few KBs. QFI produces robust optimal control signal for the current control situation using a reducing procedure and compression of redundant information in KB's of individual FCs. Process of rejection and compression of redundant information in KB's uses the laws of quantum information theory [5 – 7].

Decreasing of redundant information in KB-FC increases the robustness of control without loss of important control quality as reliability of control accuracy. As a result, a few KB-FC with QFI can be adapted to unexpected change of external environments and to uncertainty in initial information.

We introduce main ideas of quantum computation and quantum information theory [6] applied in developed QFI methods. *Quantum Fuzzy Inference* ideas are introduced. Robustness of new types of *self-organizing intelligent control systems* is demonstrated.

3 SCO-structure based on soft computing

3.1 KB of FC creation

SCO uses the chain of GAs (GA_1, GA_2, GA_3) and approximates measured or simulated data (TS) about the modeled system with desired accuracy or using real robot for it. GA_1 solves optimization problem connected with the optimal choice of number of membership functions and their shapes. GA_2 searches optimal KB with given level of rules activation. Introduction of activation level of rules allows us to sort fuzzy rules in accordance with value information and design robust KB. GA_3 refines KB by using a few criteria.

Figure 2 shows the flow chart of SCO operations on macro level and combination of several stages.

Stage 1: *Fuzzy Inference System* (FIS) *Selection.* The user makes the selection of fuzzy inference model with the featuring of the following initial parameters: Number of input and output variables; Type of fuzzy inference model (Mamdani, Sugeno, Tsukamoto, etc.); Preliminary type of MFs.

Stage 2: *Creation of linguistic values.* By using the information (that was obtained on Stage 1), GA_1 optimizes membership functions number and their shapes, approximating teaching signal (TS), obtained from the in-out tables, or from dynamic response of control object (real or simulated in Matlab).

Stage 3: *Creation rules.* At this stage we use the rule rating al-

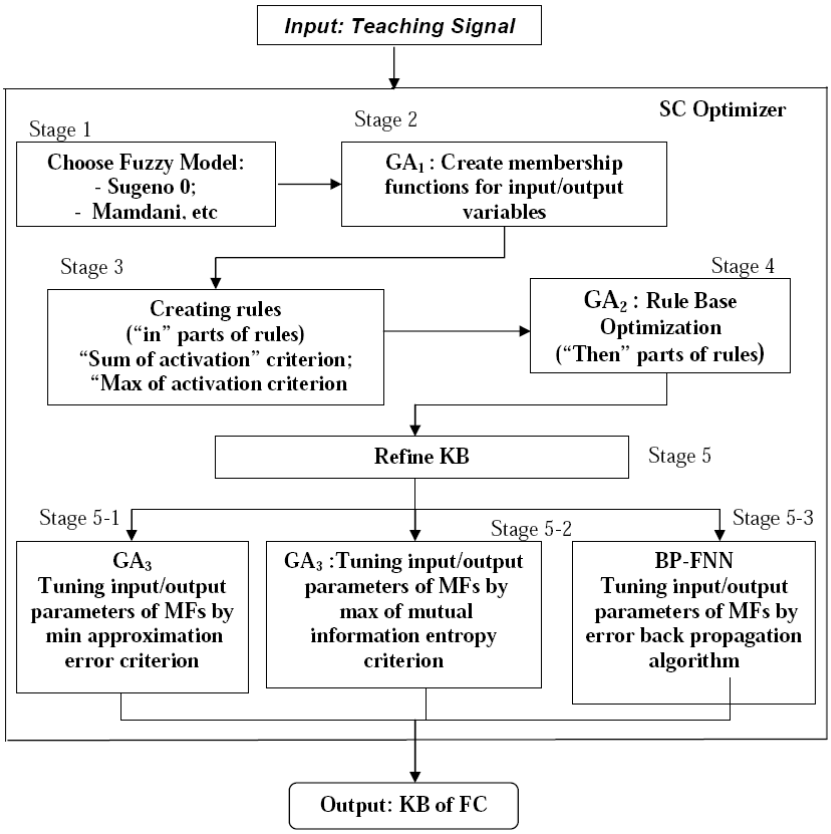


Figure 1. Flow chart of SC Optimizer

gorithm for selection of certain number of selected rules prior to the selection of the index of the output membership function corresponding to the rules. For this case two criteria are selected based on a rule's activation parameter called as a "manual threshold level" (TL). This parameter is given by a user (or it can be introduced automatically).

Stage 4: *Rule base optimization.* GA_2 optimizes the rule base obtained on the Stage 3, using the fuzzy model obtained on Stage 1, optimal linguistic variables, obtained on Stage 2, and the same TS as it was used on Stage 1. Rule base optimization can be performed by using mathematical model, or by using distance connection to real control object.

Stage 5: *Refine KB.* On this stage, the structure of KB is already specified and close to global optimum. In order to reach the optimal structure, a few methods can be used. First method is based on GA_3 with fitness function as minimum of approximation error, and in this case KB refining is similar to classical derivative based optimization procedures (like error back propagation (BP) algorithm for FNN tuning). Second method is also based on GA_3 with fitness function as maximum of mutual information entropy. Third method is realized as pure error back propagation (BP) algorithm. BP algorithm may provide further improvement of output after genetic optimization. As output results of the Stages 3, 4 and 5, we have a set of KB corresponding to chosen KB optimization criteria.

3.2 Remote rule base optimization

Remote KB optimization is performed on the fourth stage of designing FC (Fig. 1). The implementation of the physical environment connection intends to use additional equipment for the data transfer, such as radio channel, Bluetooth, WiFi or a cable connection, such as USB. Exchange of information between the management system and the SCO intended to form a KB (Fig. 2).

The control system reads the sensors and sends data to a computer for further processing. By taking input values, SCO evaluates previous decision (KB-FC) and performs fuzzy inference to check the following

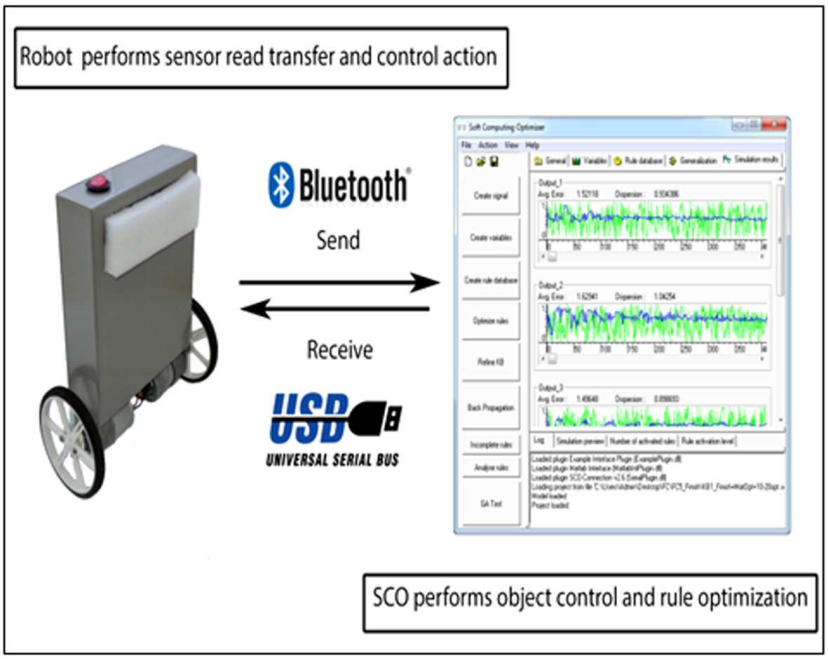


Figure 2. Remote rule base optimization scheme

solutions (KB-FC). The result of the fuzzy inference is sent to the remote device. Thereafter, the control system by processing the input values generates control action.

Synchronization of SCO and control systems is based on the remote device (robot). To this end, a special program (firmware) is developed.

Connection profile uses the serial port. Transmission rate in this case is 115,200 bits / sec. During operation, floats in symbolic form are passing via COM-port. Connection to SCO uses designed plug-in. Before establishing a connection to the SCO, COM port number and the check time of one solution (the number of cycles of the system to test solution) are selected.

4 QFI-structure based on quantum computing

For design of QFI based on a few KBs it is needed to apply the additional operations to partial KBs outputs that draw and aggregate the valuable information from different KBs. Soft computing tool does not contain corresponding necessary operations [8].

The necessary unitary reversible operations are called as *superposition*, *entanglement* (quantum correlation) and *interference* that physically are operators of quantum computing in information processing.

We introduce briefly the particularities of quantum computing and quantum information theory that are used in the quantum block QFI (Fig. 3) supporting a self-organizing capability of FC in robust intelligent control system (ICS).

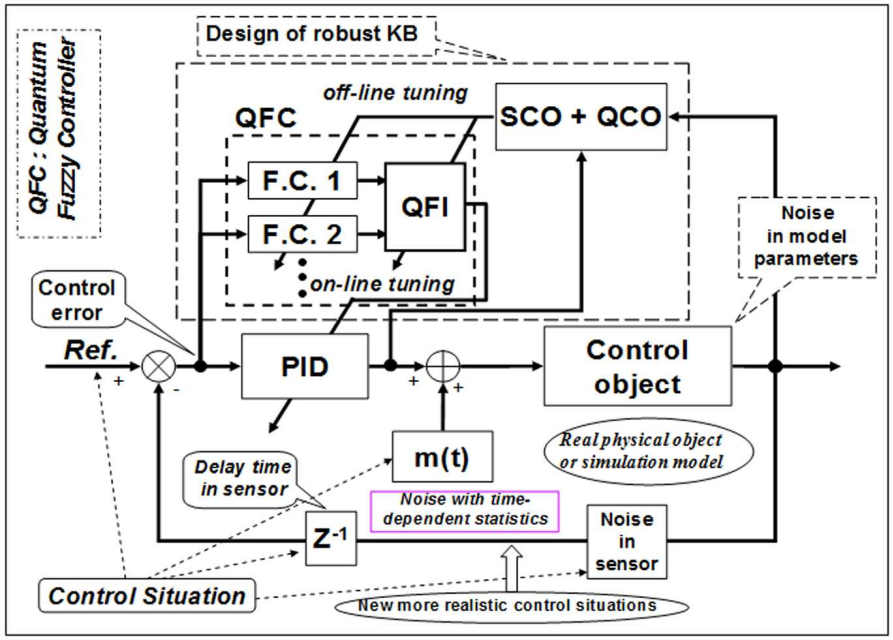


Figure 3. *Structure of robust* ICS *based on* QFI

4.1 Quantum computing

In Hilbert space the superposition of classical states $\left(c_1^{(1)}|0\rangle + c_2^{(1)}|1\rangle\right)$ called quantum bit (qubit) means that "*False*" and "*True*" are jointed in one state with different probability amplitudes, $c_i^1,\ i = 1, 2,\ \left(c_1^1\right)^2 + \left(c_1^1\right)^2 = 1$. If the Hadamard transform $H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ is independently applied to different classical states, then a tensor product of superposition states is the result:

$$|\psi\rangle = H^{\otimes n}|False\rangle = \frac{1}{\sqrt{2^n}} \otimes_{i=1}^{n} \left(|False\rangle + |True\rangle\right). \qquad (1)$$

The fundamental result of quantum computation stays that all of the computation can be embedded in a circuit, the nodes of which are the universal gates.

These gates offer an expansion of unitary operator U that evolves the system in order to perform some computation. Thus, naturally two problems are discussed: (i) Given a set of functional points $S = \{(x, y)\}$ find the operator U such that $y = U \cdot x$; (ii) Given a problem, find the quantum circuit that solves it. Algorithms for solving these problems may be implemented in a hardware quantum gate or in software as computer programs running on a classical computer.

It is shown that in quantum computing the construction of a universal quantum simulator based on classical effective simulation is possible [3, 6, 7].

In the general form, the model of quantum algorithm computing comprises the following five stages:

- preparation of the initial state $|\psi_{out}\rangle$ (classical or quantum);
- execution of the Hadamard transform for the initial state in order to prepare the superposition state;
- application of the entangled operator or the quantum correlation operator (quantum oracle) to the superposition state;
- application of the interference operator;

- application of the measurement operator to the result of quantum computing $|\psi_{out}\rangle$.

Hence, a quantum gate approach can be used in a global optimization of KB structures of ICSs that are based on quantum computing, on a quantum genetic search and quantum learning algorithms [8].

4.2 Quantum information resources in QFI algorithm

Figure 4 shows the algorithm for coding, searching and extracting the valuable information from two KBs of fuzzy PID controllers designed by SCO.

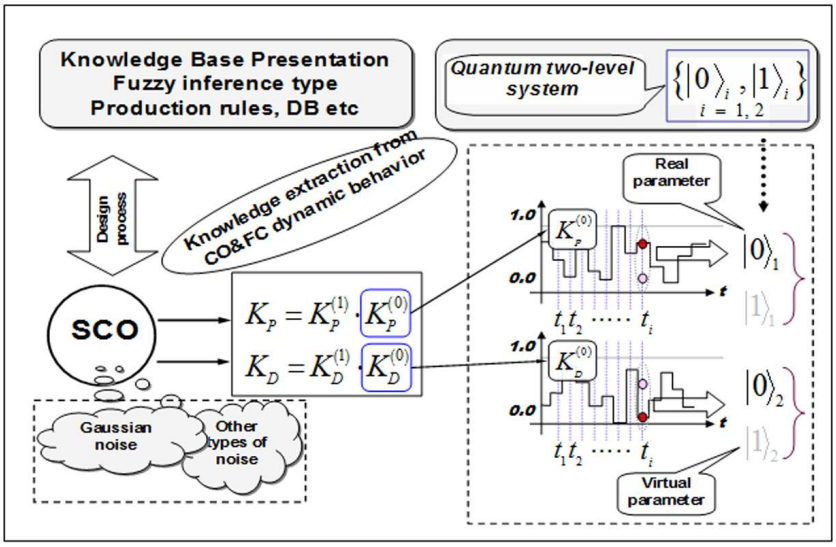


Figure 4. *Example of information extraction in* QFI

Thus, in the quantum algorithm for QFI (Fig. 5) the following actions are realized [5]:

- The results of fuzzy inference are processed for each independent FC;

- Based on the methods of quantum information theory, valuable quantum information hidden in independent (individual) knowledge bases is extracted;
- In on-line, the generalized output robust control signal is designed in all sets of knowledge bases of the fuzzy controller.
- In this case, the output signal of QFI in on-line is an optimal signal of control of the variation of the gains of the PID controller, which involves the necessary (best) qualitative characteristics of the output control signals of each of the fuzzy controllers, thus implementing the self-organization principle.

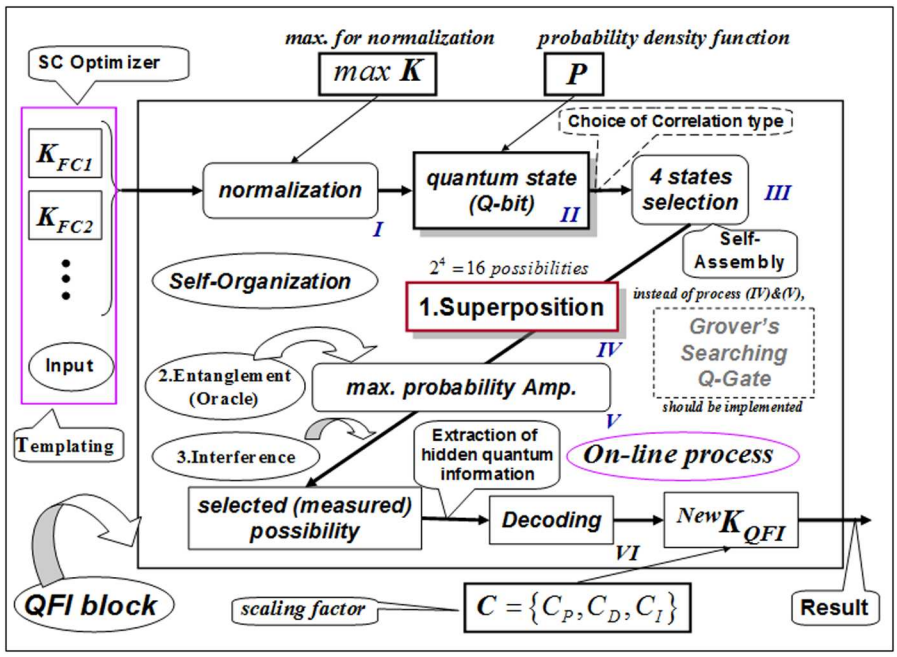


Figure 5. *The structure of* QFI *gate*

Therefore, the domain of efficient functioning of the structure of the intelligent control system can be essentially extended by including

robustness, which is a very important characteristic of control quality.

The robustness of the control signal is the background for maintaining the reliability and accuracy of control under uncertainty conditions of information or a weakly formalized description of functioning conditions and/or control goals.

QFI model based on physical laws of quantum information theory, for computing use unitary invertible (quantum) operators and they have the following names: *superposition*, *quantum correlation* (entangled operators), and *interference*. The forth operator, measurement of result quantum computation is irreversible.

Optimal drawing process of valuable information from a few KBs that are designed by soft computing is based on the following four facts from quantum information theory [4]: (i) effective quantum data compression; (ii) splitting of classical and quantum parts of information in quantum state; (iii) total correlations in quantum state are "mixture" of classical and quantum correlations; and (iv) exiting of hidden (locking) classical correlation in quantum state [6, 9].

This quantum control algorithm uses these four Facts from quantum information theory: (i) compression of classical information by coding in computational basis $\{|0\rangle, |1\rangle\}$ and forming the quantum correlation between different computational bases (Fact 1); (ii) separating and splitting total information and correlations on "classical" and "quantum" parts using Hadamard transform (Facts 2 and 3); (iii) extract unlocking information and residual redundant information by measuring the classical correlation in quantum state (Fact 4) using criteria of maximal corresponding amplitude probability.

These facts are the informational resources of QFI background. Using these facts it is possible to extract an additional amount of quantum value information from smart KBs produced by SCO for design a *wise* control using compression and rejection procedures of the redundant information in a classical control signal.

Below we discuss the application of this quantum control algorithm in QFI structure.

4.3 Remote quantum base optimization

As the adjustable parameter scaling factor is used in remote quantum base optimization. Scaling factor is used in the final step of forming the gain of PID (Fig. 5).

During operation, floats in symbolic form are passed via COM-port. The control system reads the sensors and sends them to a computer for further processing. By taking the input values, the GA evaluates the previous decision, and carries a quantum fuzzy inference to check the following solutions. The result of the fuzzy inference is sent to the remote device. Thereafter, the control system by processing the input values generates control action. Connection to QFI is developed through a plug-in.

Before establishing a connection to the SCO, COM port number and the check time of one solution (the number of cycles of the system to test solution) are selected (Fig. 6).

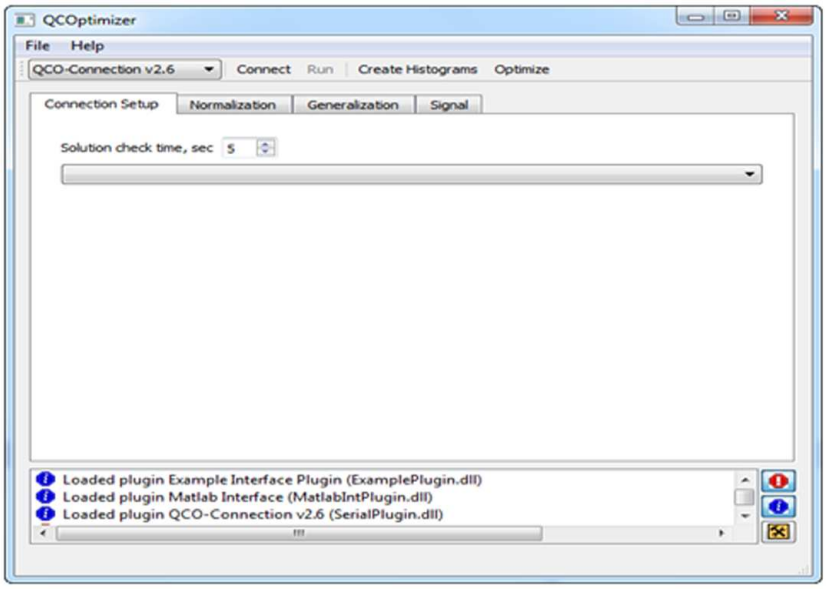


Figure 6. *Remote connection plug-in for QCOptimizer*

5 KB-self-organization of FC's based on QFI

5.1 Robust FC design toolkit

The kernel of the abovementioned FC design toolkit is a so-called SCO implementing advanced soft computing ideas. SCO is considered as a new flexible tool for design of optimal structure and robust KBs of FC based on a chain of genetic algorithms (GAs) with information-thermodynamic criteria for KB optimization and advanced error back-propagation algorithm for KB refinement [2]. Input to SCO can be some measured or simulated data (called as 'teaching signal" (TS)) about the modelling system. For TS design (or for GA fitness evaluation) we used stochastic simulation system based on the control object model. More detailed description of SCO is given in [1, 2]. Below we discuss the application of this algorithm in QFI structure.

Figure 3 illustrates as an example the structure and main ideas of self-organized control system consisting of two FC's coupling in one QFI chain that supplies a self-organizing capability. According to the described above algorithm the input to the QFI gate is considered according to (1) as a superposed quantum state $K_1(t) \otimes K_2(t)$, where $K_{1,2}(t)$ are the outputs from fuzzy controllers FC1 and FC2 designed by SCO (see Fig. 4) for the given control task in different control situations (for example, in the presence of different stochastic noises).

The algorithm of superposition calculation is presented in Fig. 7 and described in details in [4, 5].

We discuss for simplicity the situation in which an arbitrary amount of correlation is unlocked with a one-way message. Let us consider the communication process between two KBs as communication between two players A and B (see Figs 4 and 7) and let $d = 2^n$. According to the law of quantum mechanics, initially we must prepare a quantum state description by density matrix ρ from two classical states (KB_1 and KB_2).

The initial state ρ is shared between subsystems held by A (KB_1)

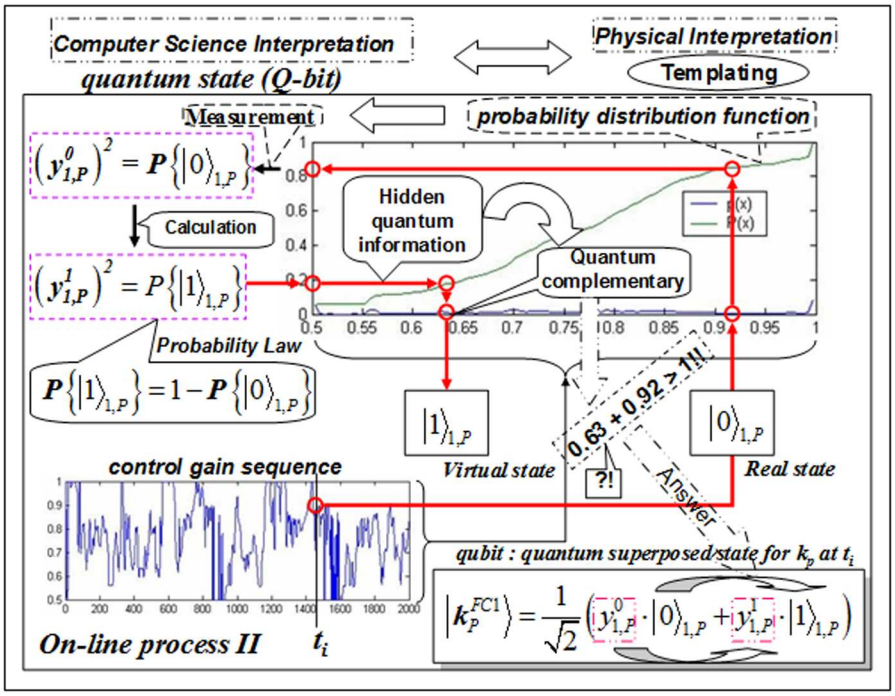


Figure 7. The algorithm of superposition calculation

and B (KB_2), with respective dimensions d,

$$\rho = \frac{1}{2d} \sum_{k=0}^{d-1} \sum_{t=0}^{1} \left(|k\rangle \langle k| \otimes |t\rangle \langle t| \right)_A \otimes \left(U_t |k\rangle \langle k| U_t^{\dagger} \right)_B . \qquad (2)$$

Here $U_0 = I$ and U_1 change the computational basis to a conjugate basis $|\langle i \mid U_1 |k\rangle| = 1/\sqrt{d} \quad \forall\, i, k$.

In this case, B chooses $|k\rangle$ randomly from d states in two possible random bases, while A has complete knowledge on his state. The state (2) can arise from the following scenario. A picks a random ρ'-bit string k and sends B $|k\rangle$ or $H^{\otimes n} |k\rangle$ depending on whether the random bit $t = 0$ or 1. A can send t to B to unlock the correlation later. Experimentally, Hadamard transform, H and measurement

on single qubits are sufficient to prepare the state (2), and the latter extracts the unlocked correlation in ρ'. The initial correlation is small, i.e. $I_{Cl}^{(l)}(\rho) = \frac{1}{2}\log d$. The final amount of information after the complete measurement M_A in one-way communication is ad hoc, $I_{Cl}(\rho') = I_{Cl}^{(l)}(\rho) = \log d + 1$, i.e., the amount of *accessible information increases*. This phenomenon is impossible classically.

However, states exhibiting this behaviour *need not to be entangled* and the corresponding communication can be organized using Hadamard transformation [9]. Therefore, using the Hadamard transformation and a new type of quantum correlation as the communication between a few KB's it is possible to increase initial information by unconventional quantum correlation (as the quantum cognitive process of a value hidden information extraction in on-line, see, e.g. Fig. 4).

In the present report we consider a simplified case of QFI, when with the Hadamard transformation there is organized an unlocked correlation in superposition of two KBs; instead of the difficultly defined entanglement operation an equivalent quantum oracle is modeled that can estimate an "*intelligent state*" with the maximum of amplitude probability in corresponding superposition of classical states (minimum entropy principle relative to extracted quantum knowledge [5]).

Interference operator extracts this maximum of amplitude probability with a classical measurement.

Figure 8 shows the structure of Quantum Computing Optimizer of robust KB-FC based on QFI [4].

The usage of the described QFI model to control non-linear dynamical systems with local and global instabilities is described below.

6 Benchmark's simulation

It is demonstrated that FCs prepared to maintain control object in the prescribed conditions often fail to control when such conditions are dramatically changed. We propose the solution of such kind of problems by introducing a quantum generalization of strategies in fuzzy inference in on-line from a set of pre-defined fuzzy controllers by new

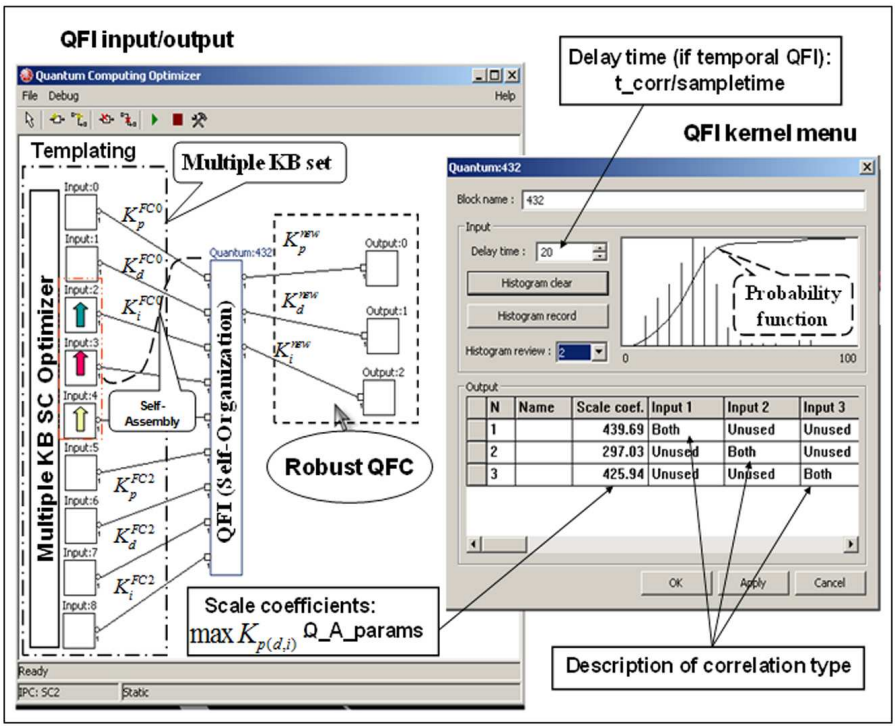


Figure 8. QFI-*process by using* QC Optimizer (QFI *kernel*)

QFI based systems. The latter is a new quantum algorithm in quantum computation without entanglement. Two Benchmarks are considered: robust control of locally and globally unstable control objects.

6.1 Benchmark 1: Globally unstable control object simulation

"Cart – pole" control object is a non-linear dissipative system. This is a typical task of control theory. The effective solution of this task demonstrates quality of control system. Task of control is the stability of inverted pendulum in vertical position. The motion of the dynamic

system "cart – pole" is described by the following equations

$$\ddot{\theta} = \frac{g\sin\theta + \cos\theta\left(\frac{u+\xi(t)+a_1\dot{z}+a_3 z - ml\dot{\theta}^2\sin\theta}{m_c+m}\right) - k\dot{\theta}}{l\left(\frac{4}{3} - \frac{m\cos^2\theta}{m_c+m}\right)} \tag{3}$$

$$\ddot{z} = \frac{u + \xi(t) - a_1\dot{z} - a_2 z + ml(\dot{\theta}^2\sin\theta - \ddot{\theta}\cos\theta)}{m_c + m},$$

where θ is the pendulum deviation angle (degrees); z is the movement of the cart (m); g is the acceleration of gravity (9.8 m/s^2); m_c is the pendulum mass (kg); l is the pendulum half-length (m); $\xi(t)$ is the stochastic excitation; and u is the control force acting on the cart (N). The equations for the entropy production rate in the control object and the PID controller have the following form, respectively:

$$\frac{d}{dt}S_\theta = \frac{k\dot{\theta}^2 + \frac{ml\dot{\theta}^3\sin 2\theta}{m_c+m}}{l(\frac{4}{3} - \frac{m\cos^2\theta}{m_c+m})};\ \frac{d}{dt}S_z = a_1\dot{z}^2;\ \frac{d}{dt}S_u = k_d\dot{e}^2 \tag{4}$$

The following parameter values are determined: $m_c = 1$; $m = 0.1$; $l = 0.54k = 0.4$; $a_1 = 0.1$; $a_2 = 5$; and the initial position $\left[\theta_0; \dot{\theta}_0; z_0; \dot{z}_0\right] = [10; 0.1; 0; 0]$ (the value of the pendulum deviation angle is given in degrees); the constraint on the control force is $-0.5 < u < 5.0$.

The specific feature of control problem for the given control object (3) is the application of one fuzzy PID controller for controlling the movement of the cart (with one degree of freedom), while the control object has two degrees of freedom.

The control goal is that the pendulum deviation angle (second generalized coordinate) reaches the given value via the implicit control using the other generalized coordinate and corresponding essentially nonlinear cross-connections with the cart movement coordinate (effect of energy transmission between the generalized coordinates).

In the case of the similar initial learning conditions, the SCO with soft computing is used to design KB_1 of FC_1 for the generalized criterion of minimal mean square error:

$$\int_{t_0}^{t_{end}} \theta^2(t)\,dt + \int_{t_0}^{t_{end}} \dot{\theta}^2(t)\,dt$$

and KB_2 for FC_2 for the generalized criterion of minimal absolute error of the pendulum position:

$$\int_{t_0}^{t_{end}} \mid \theta(\tau) \mid d\tau + \int_{t_0}^{t_{end}} \left| \dot{\theta}(\tau) \right| d\tau.$$

Thus we consider the solution of the vector (multi-objective) optimization problem based on the decomposition of the KB. The Gaussian noise was used as the random signal for designing KB_1, and Rayleigh noise was used for forming KB_2 (see Fig. 9, learning situations (**S1**, **S2**), respectively).

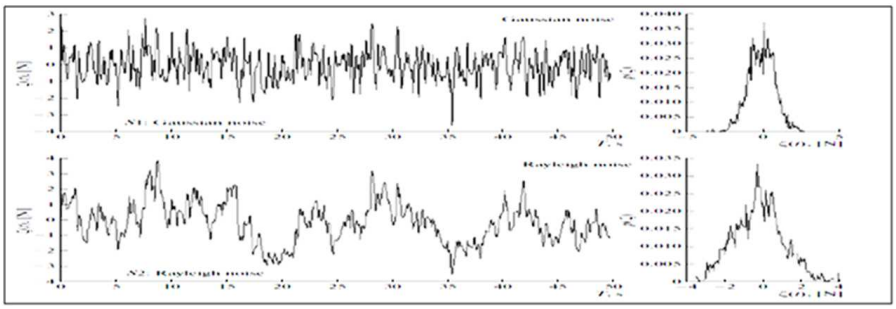


Figure 9. Random noise used in situations (**S1**, **S2**)

Physically the first criterion is equivalent to the total energy of the overturned pendulum and the second criterion characterizes the precision of the dynamic behavior of the control object.

Figure 10 shows KB_1 and KB_2 with the corresponding activated numbers of rules equal to 22 and 33 for a total number of rules of 729.

Two contingency control situations (**S3**, **S4**) were simulated; in one of them (**S3**) the new noise $\xi(t)$ was introduced, the random signal with uniform one dimensional distribution, the control error signal delay (0.03), and the noise signal in the position sensor of the pendulum (noise amplification coefficient 0.015).

Figure 11 shows the example of operation of the quantum FC for formation of the robust control signal using the proportional gain in contingency control situation **S3**. In this case, the output signals of

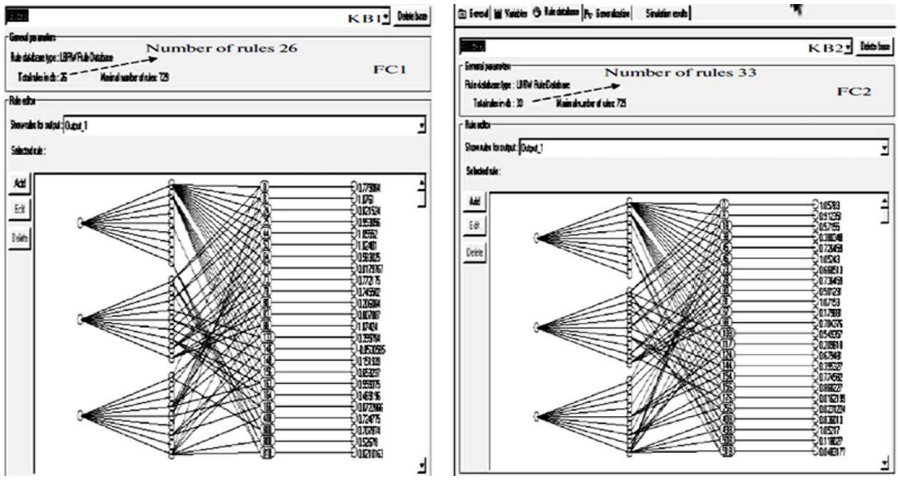


Figure 10. Form of KB_1 and KB_2 with corresponding activated production rules

KB_1 and KB_2 in the form of the response on the new control error in situation **S3** are received in the quantum FC. The output of the block of quantum FC is the new signal for on line control of the factor k_p.

Thus, the blocks of KB_1 and KB_2, and quantum FC in Fig. 3 form the block of KB self-organization in the contingency control situation.

Figure 12 shows the dynamic behavior of the studied system "cart – pole" and the control laws of the self-organized quantum controller (QFI), FC_1 and FC_2.

Remark. The following notation is used in Fig. 12 and below: $x = \theta$ is the angle of pendulum deviation from the given position; z is the cart position; the quantum FC is based on the spatial correlation.

The results of simulation (Fig. 12) demonstrate that the dynamic control object in contingency control situations (**S3**) for the control of FC_1 (FC_2) loses stability, and for the control of quantum FC the control system possesses the property of robustness and achieving the control goal is guaranteed. According to the results of simulation (Fig. 12), the required amount of control for the given criteria in contingency control situations (**S3**) for the control of FC_1 and FC_2 also is not achieved,

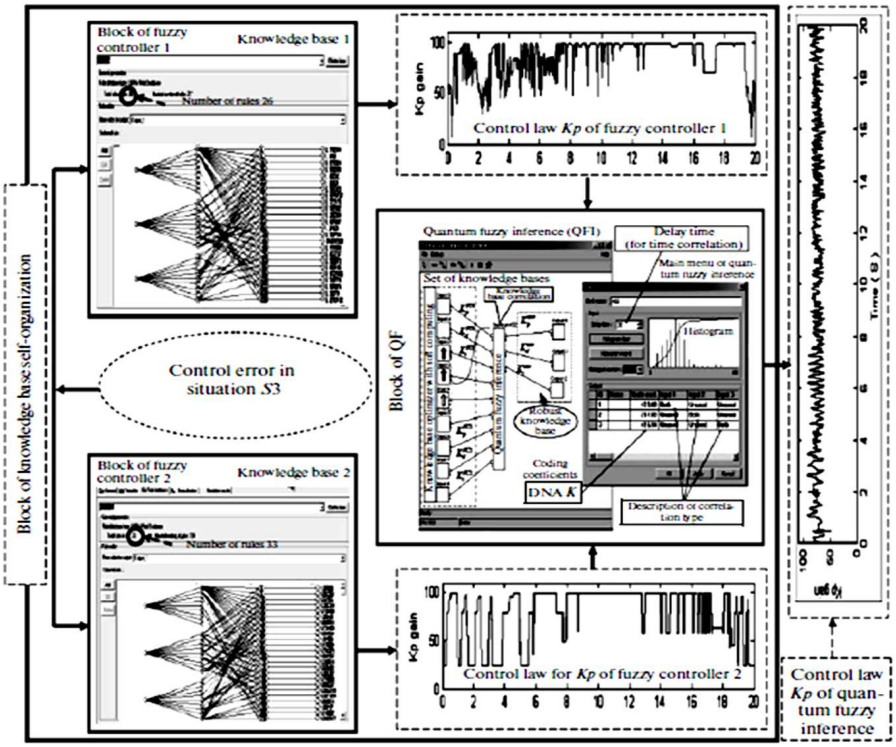


Figure 11. Example of operation of the block of KB self-organization based on QFI

while in the case of control of the quantum FC the control system possesses the required amount of control. This yields that two non robust fuzzy controllers can be used to design in on line the robust fuzzy controller using quantum self-organization; the KB of this robust FC satisfies both quality criteria.

Therefore, the decomposition of the solution to the above multi-objective optimization problem for the robust KB in the contingency control situation into partial solutions to optimization sub-problems physically can be performed in on line in the form of separate responses of the corresponding individual KBs optimized with different fixed cost

functions and control situations.

The aggregation of the obtained partial solutions in the form of the new robust KB is performed based on the quantum FC containing the mechanism of formation of the quantum correlation between the obtained partial solutions.

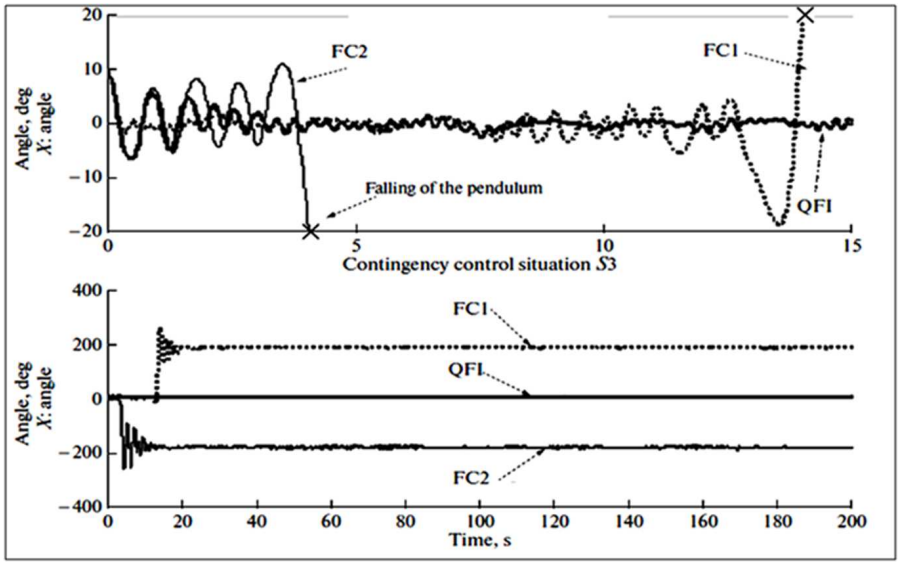


Figure 12. Dynamic motion of pole in situation **S3**

As a result, only responses of the finite number of individual KBs containing limiting admissible control laws in the given contingency situations are used.

The control laws of variation of the gains of the fuzzy PID controller formed by the new robust KB have a simpler physical realization, and as a result they possess better characteristics of individual control cost function for the contingency control situation.

For experimental testing a physical model of robot (Fig. 13) is used.

Three situations of control are tested.

The first situation images simple situation.

The second situation uses uniform noise in control channel, Gaus-

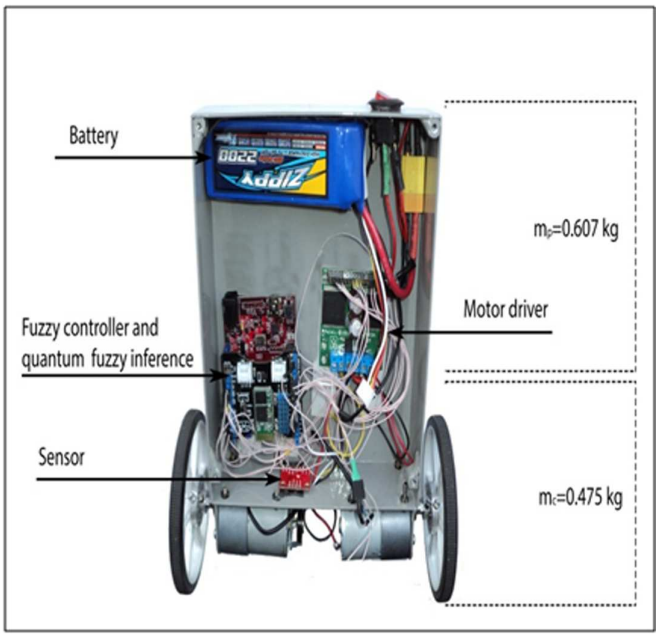


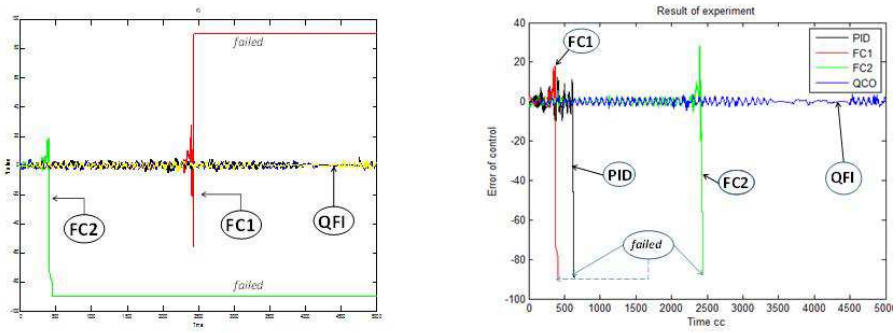
Figure 13. Mobile robot configuration

sian noise in wheel friction and delay of control action – 0.01 s. And the third situation has delay of control action equal to 0.03 s. Simulation and experimental results (for the complex situation 3) are shown in Fig. 14.

PID controller as FC_1 and FC_2 do not reach the goal in unpredicted situation.

But quantum FC based on these fuzzy controllers, are successful in unpredicted situation. For experiments and modeling we use QFI with temporal correlation, between FC_1 and FC_2.

Thus, the output signal of the quantum FC represents the on line optimal control signal for variation of the gains of the fuzzy PID controller which includes the necessary (best) qualitative characteristics of output control signals of each of the fuzzy controllers with priority and dominating component among the control quality criteria. Therefore the generalized self-organization principle [5, 8, 10 – 13] is realized.



(a) (b)

Figure 14. Control error. Unpredicted situation: (a) modeling; (b) experiment on physical model

6.2 Benchmark 2: Remote rule base optimization

To compare method of remote rule optimization on the real control object with method using Matlab simulation for optimization we created 6 KB-FC.

	TS Source	**Optimization method**	**Rules count**
FC1	Math. model	Math. modelling	125
FC2	CO (GA-PID)	Math. modelling	125
FC3	Math. model	Remote connection	125
FC4	CO (GA-PID)	Remote connection	125
FC5	Math. model	Math. modeling + Remote connection	125
FC6	CO (GA-PID)	Remote connection + Math. modeling	125

Experiment and modeling were performed in two control situations.

The first situation (S1) is typical for the control system (the initial angle equals to 1). The goal is to maintain the pendulum in equilibrium (0° angle of deflection). It should be noted that KB optimization is held

in this control situation.

The second situation is unexpected (S2). The initial angle equals to 5°. This situation characterizes the perturbation caused by external influences on CO.

Figure 15 shows the comparison of integrals of squared error for all regarded regulators in a typical situation of control:

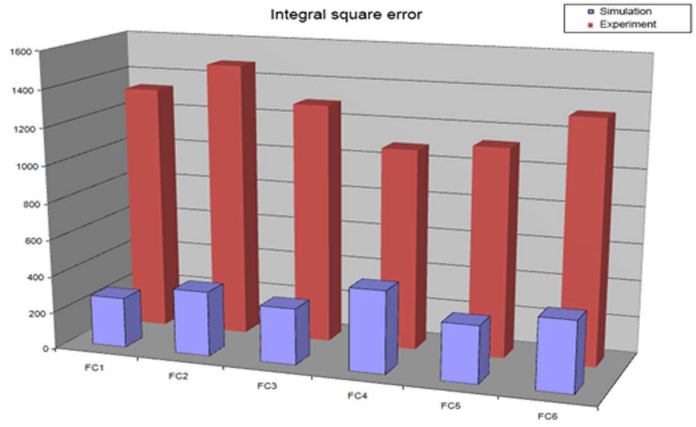


Figure 15. Integral square error. Typical situation: Simulation and experiment

The lower is integral square error level, the better controller works. Consider the results of simulation and experiment in unpredicted situation of control:

Figure 16 shows the comparison of integrals of squared error for all regarded regulators in an unpredicted situation of control.

6.3 Benchmark 3: Remote quantum base optimization

Let's compare the PID controller, fuzzy controllers FC_1 and FC_4, and QFI controllers based on different correlations: Quantum-Space (Q-S), Quantum-Time (Q-T), Quantum-Space-Time (Q-ST). These QFI controllers are optimized using remote connection.

Mathematical modeling and physical experiments took place in two control situations:

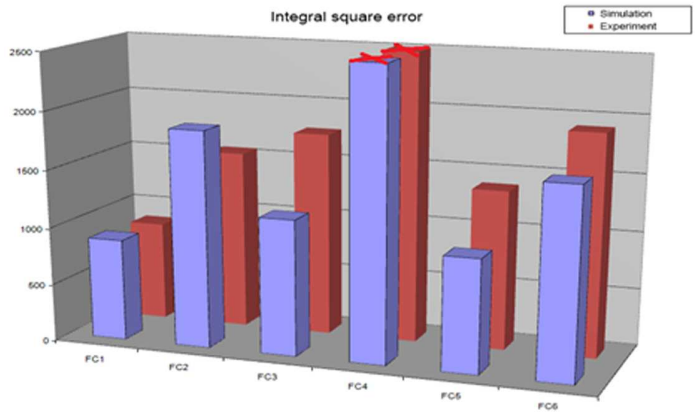


Figure 16. Integral square error. Unpredicted situation: Simulation and experiment

- in the first (typical) situation (S1), the delay of control is standard as 0.015 sec;
- in the second, unpredicted situation (S2), the delay of the control is as 0.035 sec.

From Figs 17 and 18 it can be seen that KB optimization using a remote connection with quantum optimizer can improve the quality of control in the typical and unpredicted situations.

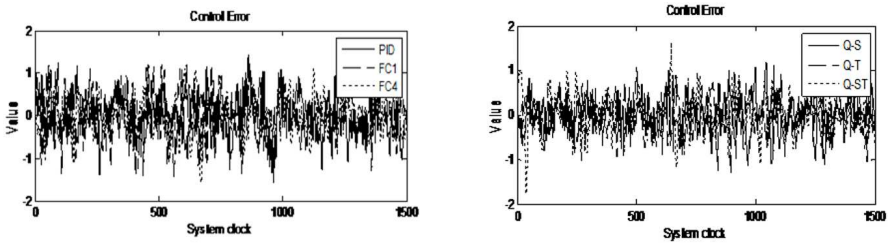


Figure 17. Control error. Typical situation of control (Experiment)

Related works. Quantum computing approaching in robot path planning, emotion design, navigation, learning, decision making was

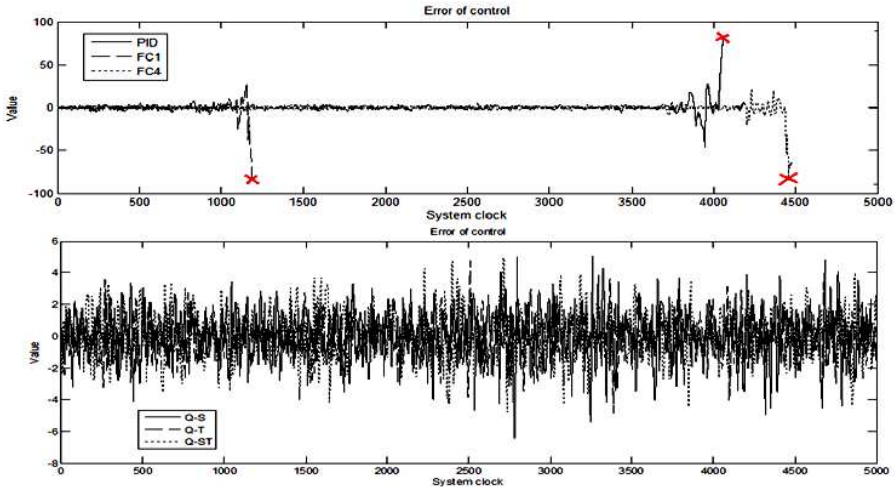


Figure 18. Control error. Unpredicted situation of control (Experiment)

applied also in [14 – 28] etc. Our approach is based on quantum self-organization of knowledge bases using responses of fuzzy controllers on unpredicted situations in on line.

7 Conclusions

The described approach opens new prospects for application of the model of quantum FC as the particular variant of the quantum self-organization algorithm in multi-objective control problems for the control object with weakly formalized structure and large dimensionality of the phase space of control parameters, application of experimental data in the form of the learning signal without development the mathematical model of the control object. These facts present a great advantage which is manifested as the possibility of design of control with required robustness in on line.

References

[1] L.V. Litvintseva, K. Takahashi, S.V. Ulyanov. *Intelligent robust control design based on new types of computation,* Note del Polo Ricerca, Università degli Studi di Milano (Polo Didattico e di Ricerca di Crema) Publ., Vol. 60, 2004; European Patent (EP) PCT 023970, 2005.

[2] L.V. Litvintseva, S.V. Ulyanov, S. S. Ulyanov. *Design of robust knowledge bases of fuzzy controllers for intelligent control of substantially nonlinear dynamic systems: II A soft computing optimizer and robustness of intelligent control systems,* J. of Computer and Systems Sciences Intern., vol. 45, No. 5, pp. 744–771; 2006.

[3] US Patent No 6,578,018B1, *System and method for control using quantum soft computing* (Inventor: S.V. Ulyanov), US Patent No 7,383,235 B1, 2003; EP PCT 1 083 520 A2, 2001.

[4] L.V. Litvintseva, S.V. Ulyanov. *Quantum fuzzy inference for knowledge base design in robust intelligent controllers,* J. of Computer and Systems Sciences Intern, vol. 46, No 6, pp. 908–961, 2007.

[5] L.V. Litvintseva, S.V. Ulyanov. *Quantum information and quantum computational intelligence: Quantum optimal control and filtering – stability, robustness, and self-organization models in nanotechnologies.* Note del Polo Ricerca, Università degli Studi di Milano (Polo Didattico e di Ricerca di Crema) Publ., vols 81&82, 2005 – 2007.

[6] M.A. Nielsen, L. Chuang. *Quantum computation and quantum information,* Cambridge Univ. Press, UK, 2000.

[7] S.V. Ulyanov, L.V. Litvintseva, I.S. Ulyanov, S.S. Ulyanov. *Quantum information and quantum computational intelligence: Quantum information, decision making and search algorithms,* Note del Polo Ricerca, Università degli Studi di Milano (Polo Didattico e di Ricerca di Crema) Publ., vols 83 – 85, Milan, 2005 – 2008.

[8] S.V. Ulyanov. *Self-organization of robust intelligent controller using quantum fuzzy inference,* Proc. of IEEE Intern. Conference ISKE'2008 (3^{rd} Intern. Conf. on Intelligent System and Knowledge Engineering), Xiamen, China, vol.1, pp. 726–732, 2008.

[9] D.P. DiVincenzo, M. Horodecki, D.W, Leung, J.A. Smolin, B.M. Terhal. *Locking classical correlation in quantum states,* Physical Review Letters, vol. 92, No 6, pp. 067902, 2004.

[10] S.V. Ulyanov, K. Takahashi, L.V. Litvintseva, T. Hagiwara. *Design of self-organized intelligent control systems based on quantum fuzzy inference: Intelligent system of systems engineering approach,* Proc. of IEEE Intern. Conf. SMC′ 2005, Hawaii, USA, vol. 4, pp. 3835–3840.

[11] S.V. Ulyanov. *Self-organized intelligent robust control based on quantum fuzzy inference,* Recent Advances in Robust Control – Novel Approaches and Design Methods / A Mueller (Ed.), Ch. 9, In Tech, 2011. pp. 187–220.

[12] S.V. Ulyanov. *Quantum soft computing in control processes design: Quantum genetic algorithms and quantum neural network approaches,* In Proc. WAC (ISSCI') 2004 (5^{th} Intern. Symp. on Soft Computing for Industry), Seville Spain, 2004, vol. 17, pp. 99–104.

[13] S. Ulyanov, A. Mishin. *Intelligent robust control of dynamic systems with partial unstable generalized coordinates based on quantum fuzzy inference,* Lecture Notes on AI (LNAI) No 7095, 2011 (MICAI 2011, I. Batyrshin and G. Sidorov (Eds.):), Part II, pp. 24–36.

[14] D. Dong, Z. - L., Chen, Z.- H. Chen, C.- B. Zhang. *Quantum mechanics helps in learning for more intelligent robots,* Chin. Phys. Lett., vol. 23, No 7, 2006, pp. 1691–1694.

[15] M. Lukac, M Perkowski. *Inductive learning of quantum behaviors,* Facta Universitatis, vol. 20, No 3, 2007, pp. 561–586.

[16] E. Kagan, Gal I Ben. *Navigation of quantum-controlled mobile robots,* Recent Advances in Mobile Robotics. Ch. 15, In Tech, 2011, pp. 311–220.

[17] A. Bannikov, S. Egerton, V. Callaghan, B.D Jonson, M. Shaukat. *Quantum computing: Non – deterministic controllers for artificial intelligent agents,* Proc. 5^{th} Intern. Wokshop Artif. Intell. Techniques for Ambient Intelligence (AITAm'10), Kuala Lumpur, Malasia, 2010.

[18] S.P Chatzis, D. Korkinof, Y Demiris. *A quantum-statistical approach toward robot learning by demonstration,* IEEE Transactions on Robotics, vol. 28, No 6, 2012, pp. 1371–1381.

[19] Kouda N., Matsui N. *An examination of qubit neural network in controlling an inverted pendulum,* Neural Processing Letters. 2005, Vol. 22, No 3, pp. 277–290.

[20] Panella M., Martinelli G. *Neurofuzzy networks with nonlinear quantum learning,* IEEE Transactions on Fuzzy Systems. 2009, Vol. 17, No 3, pp. 698–710.

[21] Chen F., Hou R., Tao G. *Adaptive controller design for faulty UAVs via quantum information technology,* Intern. J. of Adv. Robotic Systems (In Tech). 2012, Vol. 9, pp. 256–2012.

[22] Gyongyosi L., Imre S. *Quantum cellular automata controlled self-organizing networks, Cellular automata,* Innovative Modelling for Science and Engineering / Dr. A. Salcido (Ed.). InTech, 2011.

[23] Kim Y.H., Kim J.H. *Multiobjective quantum-inspired evolutionary algorithm for fuzzy path planning of mobile robot,* IEEE Congress on Evolutionary Computation (CEC 2009). 2009, pp. 1185–1192.

[24] Masood A. *A perspective on whether robot localization can be effectively simulated by quantum mechanics,* Intern. J. of Multidisciplinary Sciences and Engineering. 2012, Vol. 3, No 9, pp. 15–18.

[25] Dong D., Chen C. *Quantum robot: Structure, algorithms and applications,* Robotica. 2006, Vol. 24, No 4, pp. 513–521.

[26] Chen C., Dong D. *Quantum intelligent mobile system,* Quantum Inspired Intelligent Systems Studies in Computational Intelligence. 2008, Vol. 121, pp. 77–102.

[27] Nedjah N., Coelho L., Mourelle L. (Eds). *Quantum Inspired Intelligent Systems.* Springer Verlag, 2008.

[28] Kim S.S., Choia H.J., Kwak K., *Knowledge extraction and representation using quantum mechanics and intelligent models,* Expert Systems with Applications. 2012, Vol. 39, No 3, pp. 3572–3581.

Sergey V. Ulyanov, Andrey G. Reshetnikov, Timur A. Kerimov

Received August 23, 2013

Sergey V. Ulyanov
International University "Dubna"
Moscow, Russia
E–mail: *ulyanovsv@mail.ru*

Andrey G. Reshetnikov
International University "Dubna"
Moscow, Russia
E–mail: *reshetnikovag@pochta.ru*

Timur A. Kerimov
International University "Dubna"
Moscow, Russia
E–mail: *T.Kerimov@hotmail.com*

Computer Science Journal of Moldova, vol.**21**, no.3(63), 2013

Mental Disorder Diagnostic System Based on Logical-Combinatorial Methods of Pattern Recognition *

Anna Yankovskaya, Sergei Kitler

Abstract

The authors describe mental disorder diagnostic system based on logical-combinatorial methods of pattern recognition called as the intelligent system DIAPROD-LOG. The system is designed for diagnostics and prevention of depression. The mathematical apparatus for creation of the proposed system based on a matrix model of data and knowledge representation, as well as various kinds of regularities in data and knowledge are presented. The description of the system is given.

Keywords: intelligent system, logical-combinatorial methods of pattern recognition, diagnostic tests, intelligent instrumental software IMSLOG, depression.

1 Introduction

Creation of intelligent systems (ISs) for various semistructured areas, such as medicine, psychology, geology, etc. and development of algorithms underlying this ISs is very relevant [1, 2]. Mathematical apparatus of a number of ISs for above-mentioned problem areas is based on logical-combinatorial methods of test pattern recognition [2-4]. Currently, investigation in practical public health, viz. revealing mental and behavioral disorders is very important. However, the problem of

*This work was supported by grant from the Russian Foundation for Basic Research projects Ref. Nr. 13–07–00373a and Nr. 12–07–31109–mol_a and by grant from the Russian Humanitarian Scientific Foundation project Ref. Nr. 13–06–00709a

ISs creation for revealing regularities of various kinds, high quality and timely diagnostic and prevention of these disorders is still open. Unlike created intelligent systems [5, 6] for revealing mental and behavioral disorders in an inspected person which are based on a small number of scales and / or questionnaires, in our proposed mental disorders diagnostic system a diagnostic criteria of the international classification of diseases, tenth revision (ICD-10) [7] and 8 clinical-psychological scales and questionnaires [8-15] are used.

In Laboratory of Intelligent Systems at Tomsk State University of Architecture and Building by chief A.E. Yankovskaya have been developed ISs for revealing regularities and diagnostic and organizational-management decision-making, for example IS for revealing social-psychological factors in communicative stress conditions in learning process [16]; the IS DIOS [17] designed for express-diagnostics and intervention (correction) of organizational stress and the IS DIAPROD [18] designed for express-diagnostics and prevention of depression.

Unlike the IS express-diagnostics DIAPROD [18] using threshold and fuzzy logic for decision-making, the IS "Intelligent Decision Support System for Depression Diagnosis Based on Neuro-Fuzzy-CBR Hybrid" [5] using neural networks and fuzzy logic and the IS "Beck Depression Inventory Test Assessment Using Fuzzy Inference System" [6] using fuzzy logic, the proposed further IS DIAPROD-LOG is based on logical-combinatorial methods of test pattern recognition.

2 Basis of mathematical apparatus of creation of intelligent system DIAPROD-LOG

The mathematical apparatus of the IS DIAPROD-LOG is based on logical-combinatorial methods of test pattern recognition [3, 4]. For the data and knowledge representation in the IS DIAPROD-LOG a matrix model [4] is used.

The model includes an integer description matrix ($\mathbf{Q}$) that describes objects in the space of characteristic features $z_1, z_2, \ldots, z_m$ and an integer distinction matrix ($\mathbf{R}$) that partitions objects into equivalence

classes for each classification mechanism. A dash ("–") in the element of the matrix **Q** shows that the value of the feature is not significant to the object. We give the interval of values for each feature z_j ($j \in \{1, 2, \ldots, m\}$).

We mean under the pattern a subset of objects of knowledge base with matching values classification features.

A diagnostic test (DT) [4] is a set of features that distinguishes any pair of objects that belongs to different patterns.

A diagnostic test is called "irredundant" (dead-end [3]) if it includes an irredundant amount of features.

An irredundant unconditional diagnostic test (IUDT) is characterized by simultaneous presentation of all features of the object under investigation included in test, while decision-making.

Regularities [4] are 1) subsets of features with particular, easy-to-interpret properties that influence on the distinguishability of objects from different patterns that are stably observed for objects from the learning sample and are manifested in other objects of the same nature; 2) weight coefficients of features that characterize their individual contribution [19] to the distinguishability of objects and 3) the information weight given, unlike [20], on the subset of tests used for a final decision-making. The regularities can include constant (taking the same value for all patterns), stable (constant inside a pattern, but non-constant), non-informative (not distinguishing any pair of objects), alternative (in the sense of their inclusion in DT), dependent (in the sense of the inclusion of subsets of distinguishable pairs of objects), unessential (not included in any irredundant DT), obligatory (included in all irredundant DT), pseudo-obligatory (which are not obligatory, but included in all IUDT involved in decision-making) features and signal features, as well as all minimal and all (or part, for a large feature space) irredundant distinguishing subsets of features that are essentially minimal and irredundant DTs, respectively and tolerant to measurement (entry) errors IUDTs [21]. The weight coefficients of characteristic features calculated by different algorithms are also included in regularities [4].

There is no doubt that wider range of regularities considered provide a higher degree of accuracy while diagnostic decision-making.

We use a procedure for constructing the irredundant implication matrix ($\mathbf{U}'$) [4, 21] for revealing various kinds of regularities at construction of IUDTs.

The matrix $\mathbf{U}'$ is an integer matrix. The matrix $\mathbf{U}'$ defines distinguishability objects from different patterns (classes for each mechanism classification).

The regularities of various kinds are revealed on the matrix $\mathbf{U}'$ in order to reduce the feature space, determine the most important features. Also, all irredundant column coverings of the matrix $\mathbf{U}'$ [4, 21], defining essentially all IUDTs are determined with the use of logical-combinatorial algorithms. Then the choice of optimal subset of features is fulfilled. On the base of this subset ultimately the final decision-making is fulfilled.

3 Description of intelligent system DIAPROD-LOG

The intelligent system DIAPROD-LOG is included into the web–based complex of intelligent systems for psychological health prevention (http://psyhealth.tsuab.ru). The web–based complex is equipped with a Russian-language interface. The developed intelligent system DIAPROD-LOG is distributed. The first part of the system is designed for data collection and data and knowledge storage. This part is implemented as a web-application with the use of C#. For the purpose of data storage in the IS DIAPROD-LOG the relational database management system MySQL was chosen, since it has great flexibility, rich functionality and is free of charge.

An inspected person is offered to be tested with 8 questionnaires and scales: questionnaire A. Beck [8] including 21 features; Edinburgh post-partum depression scale [9] including 10 features; multivariate Freiburg personality inventory (FPI-B) [10] including 114 features; Tomsk questionnaire rigidity of G.V. Zalewski (TQRZ) [11] including 150 features; questionnaire of relationship of pregnant by I.V. Dobryakov [12] including 9 features; questionnaire about ways of coping R. Lazarus [13]

including 50 features; questionnaire symptom levels [14] including 90 features; questionnaire of determining the stress and social adaptation of Holmes and Rage [15] including 43 features.

Example of the survey of the IS DIAPROD-LOG is shown in fig. 1.

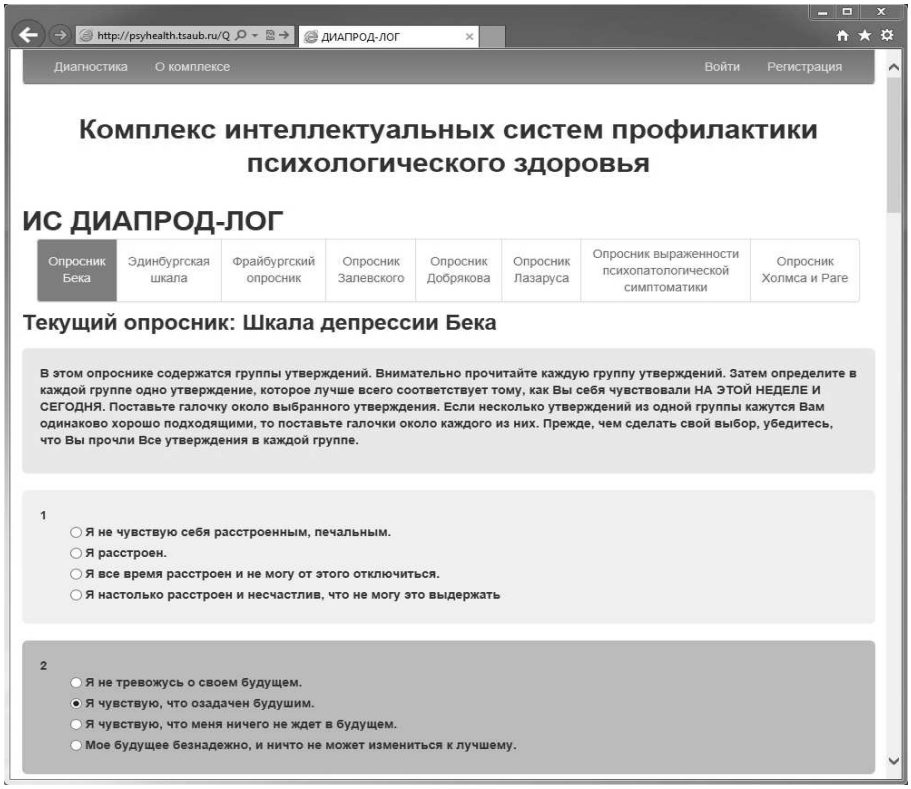


Figure 1. The survey of the IS DIAPROD-LOG

Since the inspected person confirms correctness of the entered answers, the test results are stored in the data and knowledge base.

The second part of a system designed to create matrices $\mathbf{Q}$ and $\mathbf{R}$, to construct matrix $\mathbf{U}'$, revealing different kinds of regularities, to construct diagnostic tests, decision-making and justification of decisions, is implemented as a template of intelligent instrumental software

(IIS) IMSLOG [22] including dynamically plug-ins. The template of the intelligent system DIAPROD-LOG is given in fig. 2.

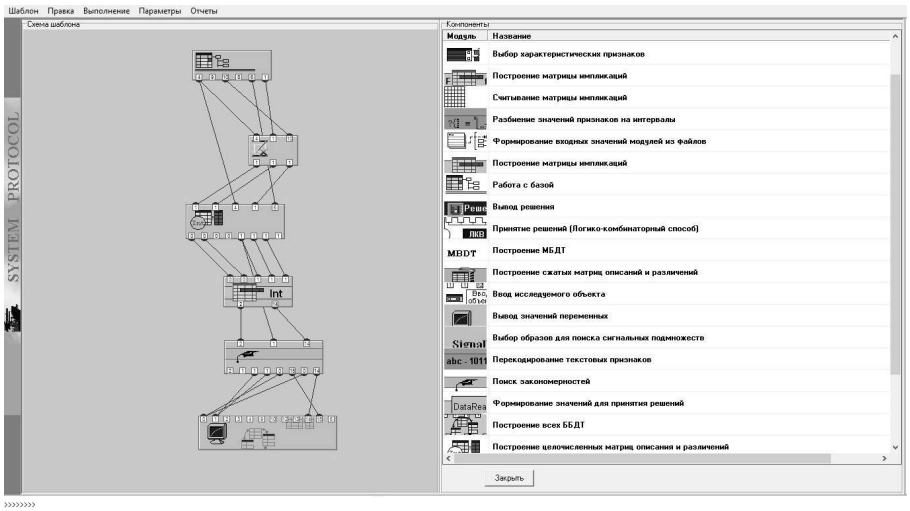


Figure 2. The template of the intelligent system DIAPROD-LOG.

The IIS IMSLOG has a module designed for data and knowledge base operation. Structure of data and knowledge base and objects of knowledge base are the input data for the module. Future selection of the necessary features for including in matrices **Q** and **R** is produced. The IS DIAPROD-LOG has characteristic features space including 28 features of the questionnaire A. Beck; the test result on Edinburgh postpartum depression scale; the test result on FPI-B; the test result on TQRZ; the test result on questionnaire of relationship of pregnant by I.V. Dobryakov; the test result on questionnaire about ways of coping R. Lazarus; the test result on questionnaire symptom levels; the test result on questionnaire of determining the stress and social adaptation of Holmes and Rage. The test results (characteristic features values) are stored in the data and knowledge base as well as classifications features values filled based on ICD-10 and highly qualified experts' knowledge in the considered problem area.

Also the IIS IMSLOG has a module designed for realizing construction of matrix $\mathbf{U}'$ using matrices $\mathbf{Q}$ and $\mathbf{R}$ with simultaneous calculating weight coefficients of characteristic features, similar to the algorithm described in [19]. In this case, the condition of tolerance to a preassigned number of measurement (entry) errors of characteristic feature values of the objects under investigation described in [19] is not implemented in this module. Then in the next module the above mentioned regularities are revealed on the basis of the matrix $\mathbf{U}'$. The next module is the construction of all irredundant column coverings of matrix $\mathbf{U}'$, defining in fact all IUDTs.

The last module fulfills the final decision-making on the diagnostic and prevention of depression in an inspected person based on voting procedure [4] on the set of tests and approaches.

4 Conclusion

The basis of the mathematical apparatus of creating intelligent system DIAPROD-LOG based on the logical-combinatorial methods of test pattern recognition, revealing various kinds of regularities, decision-making and justification decisions are suggested. The description of this system is given.

Application of the developed IS DIAPROD-LOG will allow in time diagnosing depression, making preventive decision, as well as forming the diagnostic and preventive results.

Further investigations are devoted to the intelligent system DIAPROD-LOG approbation.

We thank professor MD dean of the faculty of behavioral medicine and management of Siberian State Medical University (SSMU) and department chief of clinical psychology of SSMU Kornetov A.N. and department assistant of clinical psychology of SSMU Silaeva A.V. for consultation in questions of diagnostics and prevention of depression.

References

[1] B.A. Kobrinskiy. *A retrospective analysis of the medical expert systems*. Novosti iskustvennogo intellekta, no. 2 (2005), pp. 6–17 [in Russian].

[2] A.E. Yankovskaya. *Test recognition medical expert systems with cognitive graphics elements*. Komp'yuternaya khronika, no 8/9 (1994), pp. 61–83 [in Russian].

[3] Yu.I. Juravlev, I.B. Gurevich. *Pattern recognition and image analysis*. Artificial intelligence in 3 books, book no. 2: Models and methods: Handbook/ Edit by D.A. Pospelov (1990), Moscow: Radio i Svyaz, pp. 149–191 [in Russian].

[4] A.E. Yankovskaya. *Logical tests and cognitive graphic tools*. LAP LAMBERT Academic Publishing (2011), 92 p. [in Russian].

[5] V.E. Ekong, U.G. Inyang, E.A. Onibere. *Intelligent Decision Support System for Depression Diagnosis Based on Neuro–fuzzy–CBR Hybrid*. Modern Applied Science, vol. 6, no. 7 (2012), pp. 79–88.

[6] R.D. Ariyanti, S. Kusumadewi, I.V. Paputungan. *Beck Depression Inventory Test Assessment Using Fuzzy Inference System*. Proceeding of International Conference on Intelligent Systems Modeling and Simulation (ISMS), IEEE Computer Society, 2010, pp. 6–9.

[7] *World Health Organization, 1992. The International Classification of Diseases, Tenth Revision (ICD-10). Clinical descriptions and diagnostic guidelines, Geneva, World Health Organization.*

[8] A.T. Beck, C. Ward, M. Mendelson. *Beck Depression Inventory (BDI)*. Arch Gen Psychiatry, vol. 4, no. 6 (1961), pp. 561–571.

[9] J.L. Cox, J.M. Holden, R. Sagovsky. *Detection of Postnatal Depression: Development of the 10–item Edinburgh Postnatal Depression Scale*. British Journal of Psychiatry, vol. 150 (1987), pp. 782–786.

[10] L.I. Vansovskaya, V.K. Gajda, V.K. Gerbachevsky, et al. *Workshop on Experimental and Applied Psychology: Studies Manual / edit by A.A. Krylov*, St. Petersburg: Publishing House of St. Petersburg University (1997), 312 p. [in Russian].

[11] *Tomsk Questionnaire Rigidity of G.V. Zalewski (TQRZ) //* Siberian Psychological Journal, no. 12 (2000), pp. 129–137 [in Russian].

[12] I.V. Dobryakov. *Clinical and Psychological Methods of Determining the Type of Dominant Psychological Gestational Component.* Perinatal Psychology and Psychological Development of Children: a Collection of Conference, St. Petersburg (2001), pp. 39–48 [in Russian].

[13] T.L. Kriykova, E.V. Kuftiyak. *The Survey of Coping (Adaptation Techniques WCQ).* Journal of Practical Psychology, Moscow, no. 3, (2007), pp. 93–112 [in Russian].

[14] N.V. Tarabrina. *Practicum on psychology of post-traumatic stress.* St. Petersburg: Piter (2001), 272 p: il. [in Russian].

[15] R.V. Kupriyanov, Yu.M. Kuzmina. *Psychodiagnostics of Stress: Workshop.* The Ministry of Education and Science of the Russian Federation Kazan State Technological University, Kazan: KNRTU (2012), 212 p. [in Russian].

[16] A.E. Yankovskaya, E.A. Rogdestvenskaya. *Revealing of social-psychological factors in conditions of communicative stress in the learning process with use intelligent system.* The psychological universe of the formation of human noetic. Proceedings of International Symposium, Tomsk (1998), pp. 184–186 [in Russian].

[17] A.E. Yankovskaya, S.V. Kitler, A.V. Silaeva. *Intelligent system for diagnostics and intervention of organizational stress: its development and approbation.* Otkritoe obrazovanie, vol. 91, no. 2 (2012), pp. 61–69 [in Russian].

[18] A.E. Yankovskaya, S.V. Kitler, R.V. Ametov. *Basis for creation of intelligent system for express diagnostics and prevention of depression*. Proc. of congress on intelligent systems and information technology, vol. 2, Moscow: Fizmatlit (2011), pp. 265–272 [in Russian].

[19] A.E. Yankovskaya, A.I. Gedike. *Construction of the Implication Matrixes for Regularities Revealing in Intelligent Recognition Systems*. MEPhI-2008 Scientific Session. Collection of Papers, vol. 10, Moscow (2008), pp. 81–82 [in Russian].

[20] F.P. Krendelev, A.N. Dmitriev, Yu.I. Zhuravlev. *Comparing the Geological Structure of Foreign Depositions of Precambrian Conglomerates by Means of Discrete Mathematics*. Reports of the Sciences Academy USSR, vol. 173, no. 5 (1967), pp. 1149–1152 [in Russian].

[21] A.E. Yankovskaya, S.V. Kitler. Parallel Algorithm for Constructing k-Valued Fault-Tolerant Diagnostic Tests in Intelligent Systems. Pattern Recognition and Image Analysis, vol. 22, no 3 (2012), pp. 473–482.

[22] A.E. Yankovskaya, A.I.Gedike, R.V. Ametov, A.M. Bleikher. *IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition*. Pattern recognition and image analysis, vol. 13, no. 4 (2003), pp. 650–657.

Anna Yankovskaya[1], Sergei Kitler[2]

[1]National Research Tomsk State University;
Tomsk State University of Architecture and Building;
Tomsk State University of Control Systems and Radioelectronics
E–mail: *ayyankov@gmail.com*

[2]Tomsk State University of Control Systems and Radioelectronics
E–mail: *svkitler@gmail.com*

About IIS2013

Inga Titchiev, Svetlana Cojocaru

In the period of August, 20-23, 2013, the second edition of the International Conference ***Intelligent Information Systems*** (IIS-2013) was held at the Institute of Mathematics and Computer Science of Academy of Sciences of Moldova (ASM). It was organized by the Institute of Mathematics and Computer Science (IMCS) with the support of the Information Society Development Institute and partial financial support of Veacheslav Albu, ex-collaborator of the IMCS. This conference was dedicated to the 50-th anniversary of IMCS. Special section in memorium of the founder of researches in the field of artificial intelligence in IMCS, Iuri Pechersky, took place within IIS2013. There were 60 participants from 8 countries at this conference.

The conference objective was the discussion of a large range of subjects regarding intelligent information systems and the related ones: theoretical computer science, knowledge processing, natural language processing, decision-making, formal computing models, information technologies application in the knowledge based society, decision support systems, mono- and multiagent intelligent systems, images processing, medical diagnostic systems, intelligent business systems, human-centered computing, intelligent interfaces, etc.

There were communications presented at the conference plenary session, which wakened the interest and caused interesting debates: Jozef Gruska, "New vision and future of informatics" (Faculty of Informatics, Masaryk University, Brno, Czech Republic), Gheorghe Păun, "Some open problems about catalytic, numerical, and spiking neural P systems" (Institute of Mathematics of the Romanian Academy), Dan Tufiş, "Wiki-Translator: Multilingual Experiments for In-Domain Translations" (Artificial Intelligence Institute of the Romanian Academy), Milan Tuba, "Swarm Intelligence Algorithms Search Capabilities" (Megatrend University of Belgrade, Serbia).



In the first day a special section for young researchers was organized. At that section 8 reports of young researchers were presented. Topical problems were tackled and also the perspective ones that deal with information society development, new information technologies application, natural language processing, decision support systems, medical diagnostic systems and so on.

The ex-collaborators, colleagues and disciples of Professor Iuri Pechersky from Moldova and abroad attended the section dedicated to his memory: Ghenadii and Natalia Andrienko (Germany), Constantin Gaindric, Vsevolod Arnaut, Veacheslav Albu, Vladimir Levchenko (Moldova), Serghei Soloviev Anna Yankovskaya, Alexei Averkin, Valery Tarassov, Serghei Ulyanov (Russia) and so on.

Each of them shared his reminiscences about a person who was an outstanding, multilaterally developed and full of energy personality. Also a small gallery was organized with printed books and drawings of Iu. Pechersky, who was a catalyst for many good things in the Institute of Mathematics and Computer Science. One of the disciples of Prof.

Iuri Pechersky, Veacheslav Albu, presented the office, where professor had worked for many years, with a portrait with image of Iu. Pechersky as an acknowledgement and esteem.

As the conference result 40 papers were published and presented as the reports with authors from Republic of Moldova, Romania, Ukraine, Russia, Germany, Czech Republic, Serbia and USA.

At the end of the conference a special committee made up of leading specialists in the respective fields and present at the conference, specified a number of works by diplomas and presents: "The Young Researcher Prize" for the best young researchers' paper – Mykola Kostikov, "Decomposing morphological rules of polish for adaptive language learning", National University of Food Technologies, Kyiv, Ukraine; "The Researcher Prize" for the best paper – Verginica Barbu Mititelu, "Increasing the effectiveness of the Romanian Wordnet in NLP applications", Research Institute for Artificial Intelligence, Romanian Academy; "The Researcher Prize" for the most original presentations – Inga Titchiev, "Workflow Petri nets used in modeling of parallel architectures" (IMCS, Chisinau, Moldova), Victoria Bobicev, "Developing a question answering system" (State University of Moldova, Chisinau, Moldova). Also there were some special prizes, and namely: "The Researcher Prize" awarded to the participant, who overcame the greatest distance to come to the conference (about 5000 kilometres) – Prof. Anna Yankovskaya, Tomsk, Russia.

Abstracts of Doctor Habilitatus Thesis

Title: Small Abstract Computers
Author: Artiom Alhazov
Institute: Institute of Mathematics and Computer Science of the A.S.M.
Date of defense: August 20, 2013

Keywords: Theoretical computer science and unconventional computing, Models of computation and Turing computability, Descriptional complexity and small universal systems, P systems as parallel distributed multiset and string processing, Promoters/inhibitors and priorities, Active membranes and polarizations, Symport and antiport, Determinism and reversibility, Insertion-deletion-substitution and [hybrid] networks of evolutionary processors, Maximal and minimal parallelism and asynchronous mode.

Structure of the thesis:[1] The thesis is written in English and consists of Introduction, 6 chapters, general conclusions and recommendations, bibliography of 291 titles, 8 appendices and 219 pages of main text. The thesis includes certain numbered material: 25 figures, 9 tables, 22 definitions, 19 lemmas, 77 theorems, 47 corollaries, 9 remarks, 24 examples and 54 formulas.

Publications on the thesis topic: 120, see the author's publication webpage[2] for the list of scientific papers and links.

The concept of the unconventional computing has caught the attention of many minds, and many researchers consider it a breakthrough in theory of information processing. This is a recent and very dynamic domain of research.

The most popular motivations for the unconventional computing are those of miniaturization (as a starting point for massive parallelism or vast storage), and those of the Moore's law. We can imagine, however, other arguments to focus on the unconventional computing, without having direct applications in mind. For instance, 1) developing new algorithmic design methods for the conventional computers, 2) new perspective insights into the fundamental Physics laws, e.g., determinism, reversibility, conservation, 3) new measures of information, since this field often uses unconventional data structures, e.g., multisets, 4) new encoding methods, due to different methods of representing the information, 5) interdisciplinary research on the crossroads of Classical computing, Information theory, Number theory, Biology, Physics, etc. Unconventional computing is already a successful domain of fundamental research.

[1] `http://www.cnaa.md/en/thesis/24558/`
[2] `http://aartiom.50webs.com/`

One can consider numerous variants of models, looking for adequacy with respect to the biochemical origins of the ideas, elegance of the definitions, powerful results, or similarity with other domains of theoretical computer science.

Area of studies reflected in this habilitation thesis consists of (symbol-objects/string-objects) membrane computing and other formal computing models, mainly *distributed parallel systems rewriting multisets or strings* (e.g., networks of evolutionary processors, reversible logical elements with memory, number-conservative cellular automata, circular Post systems, insertion-deletion systems, splicing systems and ciliate gene assembly. Most of these models permit (or naturally have) parallelism and biologically inspired features. We stress, however, that the scope of the thesis is limited to studying these models as *formal computational models*, i.e., mathematical structures. The thesis consists of analysis of the corresponding domains and original research of the author. The membrane systems models considered here are: maximally parallel multiset rewriting, with/without cooperation, without or with promoters/inhibitors/priorities, deterministic or not, reversible or not; P systems with symport/antiport, with active membranes, with insertion-deletion, with ciliate operations, with energy, etc.

The main goal of the research consists in determining the computational power of restricted models. This contributes to the potential subsequent applications, answering questions about suitability of these models for the needs or possibilities of the applications.

Scientific problems solved in this thesis include: 1) Finding the computational power of a) transitional P systems with membrane creation and division, b) P systems with active membranes without polarizations, c) deterministic controlled non-cooperative P systems, d) P systems with energy. 2) Characterizing a) the class of problems polynomial-time solvable by P systems with active membranes without polarizations, b) exact power of hybrid networks of evolutionary processors with 1 node.

Theoretical significance. A number of fundamental problems of distributed parallel multiset/string processing were addressed, and the best known bounds have been proved, e.g., for the membrane systems language family and for the number of rules in maximally parallel multiset rewriting systems.

The best known results for the optimization problems considered by different authors have been established, e.g., the power of symport-3 in one membrane, the number of nodes in the hybrid networks of evolutionary processors, the number of polarizations of efficient P systems with minimal parallelism, and synchronization time of P systems.

An important characterization of rewriting systems has been obtained for

the deterministic controlled non-cooperative multiset rewriting systems.

A landscape of results was produced for the fundamental properties of multiset rewriting, such as variants of determinism, reversibility, and self-stabilizations, for multiset processing with different features (e.g., kinds of cooperation and control) working in different modes.

The computational completeness was shown for systems with very weak forms of cooperation between the elements of these systems, e.g., non-cooperative transitional P systems with membrane creation and dissolution, and P systems with polarizationless active membranes.

The optimal results were obtained for some well-studied problems, e.g., different problems for P systems with symport/antiport, different problems for P systems with polarizationless active membranes, the number of nodes in the computationally complete networks of evolutionary processors.

Impact. We only mention a few cases where further investigation by other authors emerges from the publications reflected here. A perspective research direction has been introduced, that of obligatory hybrid networks of evolutionary processors. Out of the results reflected here, those on insertion-deletion systems, have been further developed answering the original open problem. The research for P systems with active membranes computing the permanent has lead to a few subsequent breakthroughs in complexity theory of P systems.

By the completion of this thesis, DBLP[3] has shown 37 journal papers and 37 conference ones, and Google Scholar[4] has reported the authors *h-index* of 16 and *i10-index* of 30, having registered over 880 citations. His publications were presented at over 35 scientific conferences, including participation at over 20 international conferences.

Applied value of the work. One of the applications is polymorphic P systems. Its use is providing a framework where rules can dynamically change during the computation, which is important for problems of symbolic computation and computer algebra. Other applications deal with linguistics. An efficient implementation of dictionaries by membrane systems has been proposed, using membrane (tree) structure to represent the prefix tree of the dictionary. P systems were found suitable for performing inflections of words in the Romanian language. There was also proposed P systems annotating affixes of the Romanian language, also elaborating a model that accounts for complex derivation steps that may consist of multiple affixes, changing terminations and/or root alternations.

[3] `http://www.informatik.uni-trier.de/~ley/pers/hd/a/Alhazov:Artiom.html`
[4] `http://scholar.google.com/citations?sortby=pubdate&user=M8LdW5kAAAAJ`

Main scientific results. 1) A universal P system with *23 rules* has been presented in Section 3.1. 2) A detailed study of the properties of *determinism and reversibility* is given in Section 2.3. 3) Universality of transitional P systems with *membrane creation and dissolution* is proved in Section 2.5. 4) P systems with *polarizationless active membranes* are computationally complete, as presented in Section 4.2. 5) *PSPACE*-complete problems can be solved by P systems with polarizationless active membranes, as described in Section 4.5. 6) The best known results on the hybrid networks of evolutionary processors (HNEPs) are presented in Section 5.1. Specifically, *HNEPs* are universal with 7 nodes, while HNEPs with 1 node have been exactly characterized by a regular expression. 7) *Deterministic controlled non-cooperative* systems accept only finite sets and their complements; this result is described in Section 2.2. 8) It was proved that P systems with *energy* are computationally complete in the maximally parallel mode; this result is presented in Section 4.8. 9) P systems with *insertion-deletion* of a single symbol without context (with priority of deletion) are computationally complete; this result is presented in Section 5.2. 10) Besides the abovementioned theoretical results, in Chapter 6 there was described a number of *applications*, such as synchronization, polymorphism, dictionary and inflections of words in Romanian language.

Artiom ALHAZOV is a leading researcher in the Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova. He is an author of over 160 scientific publications (about 20 single-authored, also collaborating with over 50 co-authors, from 14 countries). The work of Artiom Alhazov focuses on parallel distributed multiset/string processing as an area of Theoretical Computer Science. In 2001 he graduated from the Faculty of Mathematics and Computer Science, the State University of Moldova. In 2006 he became a Ph.D. in Computer Science, Rovira i Virgili University, Tarragona. In 2013 he defended the Doctor Habilitatus Thesis in Computer Science.