# THE USE OF THE NATURAL LANGUAGE UNDERSTANDING AGENTS WITH CONCEPTUAL MODELS

Olegas Vasilecas, Algirdas Laukaitis
*Vilnius Gediminas Techical University, Saulėtekio al.11, Vilnius – 40, LT-10223 Lithuania*

Keywords: Natural language interfaces, conceptual modelling, neural network.

Abstract: In this paper AI agents for natural language interfaces in data exploration domain are presented. The experiment done with the IBM natural language toolbox has shown that the black box approach in this case leads to misclassification. Unsatisfactory results of the experiment triggered the present research aimed at improving the user interface with the natural language modality at architectural and algorithm levels. We extended traditional natural language interfaces in data exploration domain in the following direction: the use of feedforward neural network as concepts indexes in the users natural language interfaces are suggested. All presented concepts are realized as the open source project JMining Dialog.

## 1 INTRODUCTION

Corporate data environment is becoming more and more complex as the amount of information is constantly growing. Since the early 80's many efforts have been made to investigate the use of natural language for information extraction from data base management systems (DBMS). Some efforts were successful and some commercial applications emerged but the NLP techniques have not become widely used for DBMS interfaces. As mentioned by researchers in (Androutsopoulos, 1995) this is due to:

1. Graphical and menu driven interfaces achieved the level of sophistication that data analyst can do job without deep knowledge of some data queer-ing language (e.g. SQL), and on the other side NLP techniques has not been able to deliver interfaces of adequate sophistication.

2. Most research results reports on the possibility to generate only one data queering script (in most cases this was one SQL sentence) generated from one natural language sentence. They do not support complex dialog, which is the most usual case in real life when we query data analysis ex-pert.

3. Some systems are commercial products and they are close systems with difficulties in extending such systems. We think that only open source projects can bring more attention from researchers to natural language database interface systems (NLDBIS) field.

4. In available systems only system administrators are able to parameterise the system. We think that resent advances is building personal assistants in such fields like an adaptive information research from internet or personalized learning knowledge maps will help to renew researches interest in (NLDBIS) field.

To respond to these challenges a system JminingDialog (Laukaitis, 2005) that use a dialog rather than a sentence and is a constituent part of the open source information delivery web portal JMining (Laukaitis, 2005) was developed. The suggested solution presents an agent architecture consisting of a set of asynchronously operating agents. This architecture enables us to perform sophisticated data and interaction analysis without loosing the property of short respond time essential for interactive real-time operation. In the system created several well-established Java toolboxes were used. For text information pre-processing GATE (Cunningham, 2000) which is a general natural language architecture and a toolbox as well as WordNet (Miller, 1985) representing English language dictionary were applied.

The contribution of this paper is as follows: firstly, a conceptual model used in the performed experiments is described. Next the experiment with IBM natural language understanding solution WebSphere Voice Server considered as the black

Vasilecas O. and Laukaitis A. (2006).
THE USE OF THE NATURAL LANGUAGE UNDERSTANDING AGENTS WITH CONCEPTUAL MODELS.
In *Proceedings of the Eighth International Conference on Enterprise Information Systems - AIDSS*, pages 308-311
DOI: 10.5220/0002446203080311

box approach to natural language supporting systems is presented. The encountered problems stimulated the research in the uses of a hybrid neural network for natural language understanding. The main idea behind this new proposal is to simulate neural network architecture by ontological knowledge base structure.

## 2 CONCEPTUAL MODEL DRIVEN NLU UNDERSTANDING

Ambiguity and vagueness raise a lot of problems for developing information systems (IS). Business applications ambiguity and vagueness arise because at all IS lifecycle stages (analysis, design, testing etc.) natural language is an essential part of communication between people involved in business activities (policies, regulations, laws, etc.).

Conceptual data centric modelling can be an effective tool for eliminating ambiguity and vagueness from IS business applications. This can help to extend the analysts capabilities, enabling him/her to define business concepts, characteristics, behaviours, and interactions. Conceptual data-centric enterprise models are rarely built and few organizations even tried to use them with information systems and in business activities. The problem with conceptual data-centric enterprise models is that they are difficult to understand. Their abstract and generic concepts are unfamiliar to both business users and IS professionals, being removed from their local organizational contexts. We have found this in several Baltic and Scandinavian banks working with the IBM financial services data model (FSDM) (IBM, 2005), which is a domain specific model, based on the ideas of the experts from IBM banking solution centre. To boast the awareness and project-centric approach we integrated the model into the created data exploration and information extraction framework JMining (Laukaitis, 2005). The model is shown to consist of a high level strategic classification of domain classes integrated with particular business solutions (e.g. Credit Risk Analysis) and logical and physical data entity-relationship (ER) models. In JMining Dialog system the user identifies concepts by using natural language on all conceptual models levels: the 'A' level identifies nine data concepts that define the scope of the enterprise model (involved party, Products, arrangement, event, location, resource items, condition, classification, business), the 'B' level contains with business concepts hierarchies (more than 3000 concepts), the 'A/B' business solutions (integrates more than 6000 concepts with more than 50 solutions) and 'C' level – entity relationship ER diagram with about 6000 entities, relationships and attributes.

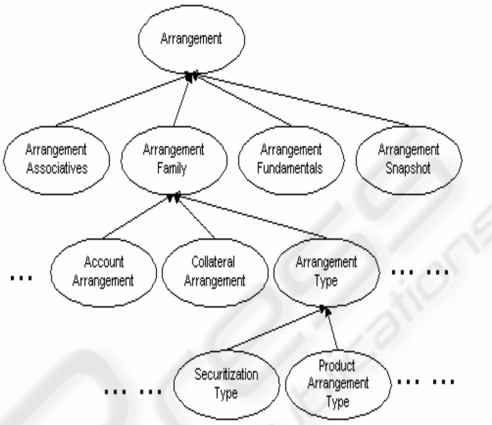

Figure 1: Part from conceptual model used to support natural language modality in data querying system Jmining.

In figure 1 we can see the small part from conceptual model. If the user brings the input, "show all arrangements with the type loan", the system activates the conceptual model graph paths with different probabilities for each concept e.g.: 1) Arrangement (0.59) -> Arrangement Family (0.42) -> Account Arrangements (0.40) -> Loan Arrangements (0.14), 2) Arrangement (0.59) -> Arrangement Family (0.42) -> Arrangements Type (0.25) -> Product Arrangements Type (0.23) etc.

As we see the user natural language input activates not just one concept but a path on conceptual graph. Then intelligent agents can act on that information e.g. agent responsible for SQL understanding can build the SQL sentences from identified databases, agent responsible for dialog handling can propose several options for user and ask to specify more accurately what the user has in mind.

## 3 NATURAL LANGUAGE UNDERSTANDING WITH IBM NLU TOOLBOX

At the beginning of the research we looked for the state-of-the-art natural language understanding (NLU) systems that can be found in the market and used as plugging to our concepts identification system. We have made primary evaluation of WebSphere Voice Server, which is a part of the IBM WebSphere software platform. From IBM presentation (IBM, 2004) it appeared that the system is primarily intended for telecommunication market. It was a challenging task to test it on more a complex system e.g. a full conceptual model for financial services. The IBM NLU system uses statistically based models, which as they claim, provide more flexibility and robustness compared with traditional grammar-based methods. Much of the algorithm is unknown becouse the product is proprietary. In the present research the black box approach was used: put the training data, compile and test the system response to the new arriving data. For statistical learning the sets of pairs including the concept and the description of the concept were provided.

The following experiment conducted with IBM NLU solution revealed some basic problems with the current state-of-the art technologies when we want to apply them beyond ordinary telephony voice applications. A group consisting of 3 students was instructed about the above data model. They queried the system with about 20 questions and tried to identify the "Involved Party" concept. The number of concepts put into IBM NLU model for learning was constantly increased. At the beginning only 9 top 'A' level concepts were considered. In this case for training data a description of these concepts were extracted from the original IBM model. At the second stage, the descriptions from child concepts were added to the training data for these 9 top parent concepts (see the second row in the table). Next the number of concepts was increased to 50 and finally 500 concepts with their descriptions were extracted and put to the IBM NLU statistical training data. Table 1 shows the results of the experiment. To detect the classification error the proportion of the correct identified concepts was used.

We were faced with a critical scalability problem. There were several instances in training when the system diverged from any reasonable acceptance level. While it was possible to make the training successful through manual intervention by adding more training data, the problem of divergence remained when the number of concepts increased up to the full conceptual model. The present research has shown that there is a lack of descriptive power for entities identification when training data include only brief descriptions of the conceptual model entities (as in IBM FDWM).

Table 1: Concepts identification experiment (CN - number of concepts for identification).

| | CN=9 | CN=50 | CN=500 |
|---|---|---|---|
| 1. IBM NLU | 0.1521 | .0405 | 0.0152 |
| 2. IBM NLU (child nodes descriptions added) | 0.3682 | .1726 | 0.0822 |
| 3. Hybrid modular FF NN (NL parsers integrated in the network structure) | 0.4590 | 0.2814 | 0.1874 |

To increase concept identification accuracy, we experimented with Separate Multi-Layer Feedforward Network (MLF) with one hidden layer. The novelty of this experiment is that there is a feedforward network representing each node (concept) in the conceptual model. To train the network unit, which represents one node, we suggested that a different dictionary be provided for each network. For parent nodes children's training data, which was used in the IBM NLU experiment, was employed. In the presented architecture each network is concentrated on identification of one entity, but each network has a connection with other networks representing different concepts.

It has been found that such "weak" connectionism between separate neural networks can increase concept identification. First the modular network was tested without symbolic pre-processing. In the training process concept maps were constructed based on the training examples. These concept maps relate each input sentence/phrase to a specific concept in the problem domain. All patterns consist of a unipolar representation of the training sentence or phrase. For example, the sentence could be: Show all my arrangements. Then the pattern for concept arrangement would be: 1 0 0 0 … 0 0 … .

It has also been found that if there is a case where there is no symbolic preprocessing there should be textual input that accurately matches the network dictionary. This was the main reason why we decided to improve the performance of the system by transforming our dictionary input into Vector Space Model VSM. For this purpose, methodology presented in (Wermter, 1995) based on WordNet (Miller, 1985) was used for additional semantic mapping. Term weighting is a well-known representation approach that transforms a term to a weight vector in text processing. For neural models, this representation plays a key role in model performance. The most common term-weighting method, is based on the bag-of-words approach, which ignores the linear ordering of words within the context and uses basic occurrence information. In addition the GATE (Cunningham, 2000) was used to extend semantic mapping of the WordNet initially used by others researches (Wermter, 1995). With GATE toolboxes some natural language processing techniques, such as tagging, parsing, and word sense disambiguation can be integrated with statistical word knowledge.

Table 1 shows the results of the experiment. Row 3 demonstrates that symbolic natural language processing combined with the connectionism paradigm can improve concepts prediction accuracy.

## REFERENCES

Androutsopoulos, I., Ritchie, G.D., Thanisch, P., 1995. Natural Language Interfaces to Databases - An Introduction. *Natural Language Engineering*, 1(1):29-81.

Androutsopoulos, I., Ritchie, G.D., Thanisch, P., 1995. Experience Using TSQL2 in a Natural Language Interface. In *J. Clifford and A. Tuzhilin, editors, Recent Advances in Tem- poral Databases - Proceedings of the International Workshop on Temporal Databases, Zurich, Switzerland, Workshops in Computing*, pages 113-132. Springer-Verlag, Berlin.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Wilks, Y., 2000.Experience of using GATE for NLP R/D. In *Proceedings of the Workshop on Using Toolsets References 200 and Architectures To Build NLP Systems at COLING-2000*, Luxembourg.

IBM IBM Banking Data Warehouse General Information Manual. Available from the IBM corporate site http://www.ibm.com (accessed July 2005).

IBM. An Introduction to IBM Natural Language Understanding. *An IBM White Paper*. Available from the IBM corporate site http://www.ibm.com (accessed July 2004).

Laukaitis, A., Vasilecas, O., Berniunas, R., 2005. JMining - information delivery web portal architecture and open source implementation *// Edited by O. Vasilecas et al. Information Systems. Development: Advances in Theory, Practice and Education*., Springer.

Laukaitis, A., Vasilecas, O., 2005. An architecture for natural language dialog applications in data exploration and presentation domain. ADBIS.

Miller, G.A., 1985. WordNet: A Dictionary Browser, *Proc. 1st Int'l Conf. Information in Data*, pp. 25-28.

Wermter, S., 1995. Hybrid Connectionist Natural Language Processing, *Neural Computing Series*, Chapman & Hall.