

Chatterbot for Education: a Study based on Formal Concept Analysis for Instructional Material Recommendation

Sílvia M. W. Moraes, Rackel M. Machado

Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul
(PUCRS)

Caixa Postal 1429 – 90619-900 – Porto Alegre – RS – Brazil

silvia.moraes@pucrs.br, rackel.machado@acad.pucrs.br

***Abstract.** This paper describes an ongoing research. We are building a chatterbot for education, in Portuguese, which uses conceptual lattices to generate answers for the students. We automatically build conceptual lattices from instructional materials about Artificial Intelligence. In this paper we describe how the lattices are generated for chatbot system.*

***Resumo.** Este artigo descreve uma pesquisa em andamento. Nós estamos construindo um chatterbot para a educação, em Português, que usa reticulados conceituais para gerar as respostas para os estudantes. Nós geramos automaticamente esses reticulados conceituais a partir de materiais instrucionais sobre Inteligência Artificial. Esse artigo nós descrevemos como os reticulados do chatterbot são gerados.*

1. Introduction

The evolution of the Internet combined with the expansion of systems for web, especially for mobile devices, has created new demands in different areas of knowledge. According Banchs and Li (2012), this is the case of chatterbots, which have gained a lot of popularity and are being considered at present "modern channel of communication between company and its consumers" (Kuligowska and Lasek, 2011) . Conceptually, chatterbots are dialog systems that aim to communication in natural language, with users in order to assist them in any way (Jurafsky and Martin, 2005; Al-Zubaide and Issa, 2011). Chatterbots, also called chatbots or conversational agents, are designed to emulate usual human responses during a conversation (Kuligowska and Lasek, 2011).

In education, especially in e-learning, the chatbots have received much attention. The reasons for this interest are focused on their characteristics. The chatbots provide a more natural and practical interface to the student. They are able to offer a personal support, to recognize the strengths and abilities of students. In addition, they can also accelerate the learning to act as instigators of study topics, resulting in more engaged and independent learners (Kerly et al, 2009; Griol et al 2011; Moreno et al, 2015; Ramayana and Netto, 2014). The students showed more motivation when there is a more personal interaction. According to Ghose and Barua (2013). students often prefer the interaction with chatbots, even when the information is available on the web.

Despite the popularity of these systems, most chatbots use only trivial tools for natural language processing (NLP). Moreover, their knowledge bases in general are written in AIML¹, which is based on pattern matching. An AIML knowledge base is

¹ AIML means Artificial Intelligence Markup Language. AIML is an XML version that is designed to create the

easy to be defined by not requiring any computational knowledge to its creation.

However, it must contain a lot of question-answers patterns for a similar dialogue to natural. This results in a lot of manual effort and makes the inflexible system. In order to reduce this effort and improve the quality of interaction conversational agent, we propose an architecture that includes treatment of user input through NLP techniques and the organization of instructional material (articles, books, web pages, ...) in an ontological structure based on Formal Analysis Concept (Stumme et al, 1995). The Formal Concept Analysis (FCA) method is applied in different areas. In computer science, the method is considered a conceptual clustering technique. This method generates groups of concepts and provides an intensional description to these groups. This description favors the traceability of the process of building ontological structures and renders the generated groups easier to interpret. In our research, we automatically generate the conceptual lattices (ontological structures) from information extracted from text-based instructional materials in Portuguese. The main objective of our study is to develop a conversational agent that aids in learning, acting as a facilitator in the access to information about study topics.

This paper is organized in five sections. Section 2 provides a brief introduction to chatterbots. Section 3 describes the Formal Concept Analysis method. Section 4 presents the proposed architecture. And in the Section 5 are our final remarks.

2. Chatterbots

The chatterbot term is a concatenation of the words: chatter (people chatting) and bot (short for robot - robot). Although the first chatterbot has arisen in the 60's, this term was first used in 1994 by Michael Mauldin to designate a player controlled by computer (Mauldin, 1994). Historically, the first chatbot was ELIZA. This system was built in 1966 by Joseph Weizenbaum to simulate a psychotherapist (Weizenbaum, 1996). Since, different chatterbots were created, being chatbot ALICE (Artificial Linguistic Internet Computer Entity) the best known. ALICE was created by Richard Wallace and was activated in 1995 (Wallace, 2003). It was the first program based on AIML. Recently, chatbots have been used in various fields. In education, they are often used in tutoring systems, question-answer and others learning system to stimulate thinking and metacognitive ability of students. They also act as pedagogical agents and partners in the learning process (Kerly et al, 2009; Griol et al 2011).

Despite the popularity of these systems, most chatterbots use trivial NLP tools. In addition, many of these systems are based on AIML (Al-Zubaide et al, 2011; Klüwer, 2011). Chatterbots look for patterns into user input and try to associate these patterns to templates (answers) from AIML basis. According Kluwer (2011), the choice by AIML is due to two reasons: development more simple and robustness to unexpected input. However, AIML bases require many patterns of question-answers for a similar dialogue to natural. This results in a lot of manual effort and makes the inflexible system. In order address these problems, Kluwer suggests optimizations ranging from system integration with other knowledge bases to use NLP techniques more elaborated. And these problems are the main motivations for our study.

3. Formal Concept Analysis

Formal Concept Analysis (FCA) is a method used for data analysis, knowledge

stimulus-response dialog systems.

representation and information management. The main idea behind the method is the duality known as the “Galois connection”, which establishes implicit relations between objects and attributes in such a way that objects can be described by their attributes and the attributes by the objects they characterize (Stumme et al., 1995).

A key element in this method is formal context, characterized by the triple (G, M, I) , where: G is the set of domain entities, called formal objects; M consists of the features of these entities, their formal attributes; and I is the binary relation of $G \times M$, called the incidence relation, which associates a formal object to its attribute (Stumme et al., 1995). In this paper, we build formal contexts based on syntactic information. We define *bigrams – texts* pairs, where the bigrams² are obtained from the complements (subject, direct and indirect object) of the verbs. Table 1 presents an example of formal context. The bigrams constitute the set of formal objects $G = \{inteligência_artificial, algoritmo_genético, algoritmo_Astar, agentes_inteligentes\}$ and the texts from instructional material constitute the set of formal attributes $M = \{t1, t2, t3, t4\}$. Set I establishes the relations between the elements in G and those in M , such as that of the formal object “algoritmo genético” and its attributes, “t2”, “t3” and “t4”.

Table 1. Example of Formal Context

Objects/Attributes	t1	t2	t3	t4
algoritmo_genetico		x	x	x
algoritmo_Astar		x	x	x
agentes_inteligentes	x	x		x
inteligencia_artificial	x	x	x	x

Based on the formal context, the formal concepts are built. A formal concept in (G, M, I) is determined by the pair (O, A) if and only if $O \subset G, A \subset M$, so that $f(O)=A$ and $h(A)=O$, where the operator f defines the properties shared by all the elements in O and the operator h determines the objects that contain all of the properties in A . The following pairs are examples of formal concepts of FCA illustrated in Figure 1: $(\{algoritmo_genetico, inteligencia_artificial, algoritmo_Astar\}, \{t2, t3, t4\})$ and $(\{agentes_inteligentes, inteligencia_artificial\}, \{t1, t2, t4\})$.

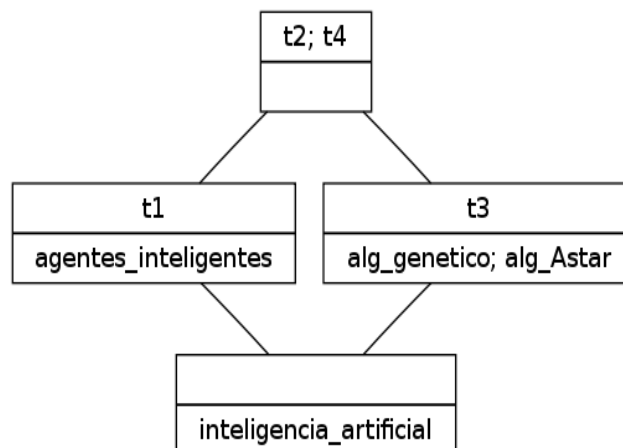


Figure 1. Example of conceptual lattices using FCA method.

2 A bigram is a sequence of two adjacent words.

4. Architecture

The main objective of our research is the development of a conversational agent that aids in learning, acting as a facilitator in the access to information about study topics. Our goal is to generate responses to student including the recommendation of instructional materials. Our architecture allows to built chatbots for any domain. In our study, we are using texts in Portuguese about Artificial Intelligence (AI). These texts are already classified by topic (eg: agents, neural networks, knowledge representation, etc.). From these texts, we automatically generate conceptual lattices using FCA method (Cimiano, 2006). For this, the texts, initially, are preprocessed. We transform acronyms and abbreviations, we apply lemmatization³ (morphological normalization best suited for Portuguese), we include Part-Of-Speech (POS) annotation, we remove accents and generate n-grams⁴. Then, we apply a feature selection technique to build the bag-of-words⁵ (*bow*). We are testing simple techniques such as ranking and relevance for this purpose. We generate one *bow* to each topic of AI. After, we generate a formal context where the objects are items from the *bow* and attributes, texts. Defined the formal contexts, we generate a conceptual lattice for each topic of AI (Figure 2). These lattices are used to compose answers to students.

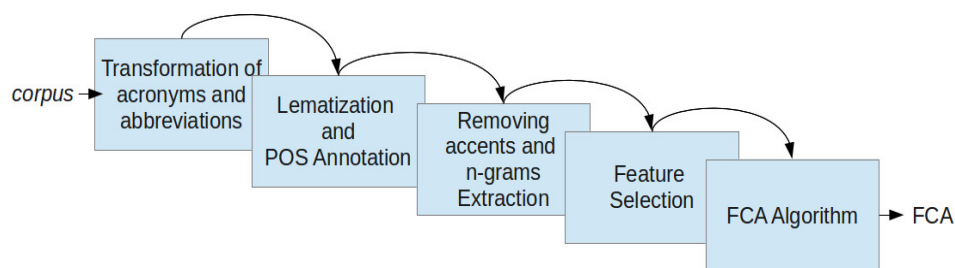


Figure 2. Construction of conceptual lattices from instructional material.

During the interaction of chatbot with the student, the inputs are preprocessed in the same way that the texts were. We also replace some words from user input by synonyms. After, these inputs are matching to pattern of a AIML base. We use a AIML base manually built. In this base, the answers describe definitions about AI terms. Through an interpreter for AIML (PyAIML), the system looks for a correspondent template to user input. Then, the system search, in the appropriate lattice, the instructional material that complements the response and it include, in the output for student, a link to this material.

5. Acknowledgment

This research has the financial support of PUCRS (EDITAL N. 01/2015 – Programa de Apoio à Atuação de Professores Horistas em Atividades de Pesquisa na PUCRS).

6. Final Remarks

We present in this paper an architecture for educational chatbot based on FCA. Our approach is promising to explore the inherent characteristics of the FCA and because we

³ Lemmatization allows to represent the different inflected forms of a word as single term (lemma of word).

⁴ An n-gram is a contiguous sequence of n items from a given sequence of text or speech. In this case, items are words.

⁵ In bag-of-words model, a sentence is represented as a bag of its words, disregarding grammar and even word order but keeping multiplicity.

yet do not find papers that use FCA as a method to recommend instructional material automatically. After we complete the implementation of this architecture, we plan to test it in undergraduate students as well as extend the study to other areas of knowledge.

References

- Al-Zubaide, Hadeel and Issa, Ayman A. (2011) "OnBot: Ontology based ChatBot". Fourth International Symposium on Innovation in Information & Communication Technology, IEEE, p. 7-12.
- Banchs, Rafael E. and Li, Haizhou. (2012) "IRIS: a Chat-oriented Dialogue System based on the Vector Space Model". Proceedings of the 50th Annual Meeting of the Association for Computational Linguistic, Korea, p. 37-42 .
- Cimiano, P. (2006) *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, USA: Springer-Verlag, 347p.
- Glose, Supratip and Barua, Jagat Joyti. (2013) "Toward the implementation of a Topic specific Dialogue based Natural Language Chatbot as an Undergraduate Advisor", IEEE.
- Griol, David; Garcia-Herreno, Jesús and Molina, José M. (2011) "The EduAgent Platform: Intelligent Conversational Agents for E-learning Applications". *Advances in Intelligent and Soft Computing*, Volume 92, Springer, p. 117-124.
- Jurafsky, Daniel and Martin, James H. (2005) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.
- Kerly, Alice; Ellis, Richard and Bull, Susan. (2009) "Conversational Agents in E-Learning". *Applications and Innovations in Intelligent Systems XVI*, Springer, p. 169-182.
- Klüwer, Tina. (2011) "From Chatbots to Dialog Systems." In *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*, ed. Diana Perez-Marin and Ismael Pascual-Nieto, p. 1-22.
- Kuligowska, K. and Lasek, M. (2011) "Virtual assistants support customer relations and business processes". *Information Management*, Gdańsk University Press.
- Mauldin, Michael L. (1994) "Chatterbots, TinyMuds, and the Turing Test Entering the Loebner Prize Competition, AAI'94, Proceedings..., p. 16-21.
- Moreno, F.C., Manfio, E.R., Barbosa, C.R and Brancher, J.D. (2015) *Tical: Chatbot sobre o Atlas Linguístico do Brasil no WhatsApp*, CBIE.
- Ramayana Junior, A.M and Netto, J. F. M (2014) "Um Chatterbot Educacional Baseado em EmotionML", CBIE.
- Stumme, G., Darmstadt, T. H. and Mathematik, F. (1995). Exploration tools in formal concept analysis, In *Proceedings of Studies in Classification, Data Analysis, and Knowledge Organization 8*, Springer,31-44.
- Wallace, Richard. (2003) "The Anatomy of ALICE", <http://www.alicebot.org>.
- Weizenbaum, Joseph. (1996) "ELIZA – a computer program for the study of natural language communication between man and machine". *Communication of the ACM*, 9(1), New York, USA, p. 36-45.