Intelligent Chatter Bot for Regulation Search

María Daniela López De Luise
CI2S Lab,
Ciudad Autónoma de Buenos Aires, Argentina
daniela_ldl@ieee.org

Andrés Pascal, Rafael Malgor
Universidad Tecnológica Nacional, FRCU,
andrespascal2003@yahoo.com.ar,
rafaelmalgor@gmail.com,

Ben Saad, Pablo Pescio,
Claudia Álvarez, Joaquín Díaz
Science and Technology Dep., UADER
Concepción del Uruguay, Entre Ríos, Argentina
ben.saad@gmail.com, pdpescio@yahoo.com.ar,
claudialvarez2000@yahoo.com.ar, mjoaquind@gmail.com

Patricio Carrilero
CI2S Lab, Ciudad Autónoma de Buenos Aires, Argentina

Abstract— *This communication presents a functional prototype implementing a linguistic model focused on regulations in Spanish. Its global architecture, the reasoning model and short statistics are provided for the prototype named PTAH. It mainly has a conversational robot linked to an Expert System by a module with many intelligent linguistic filters, implementing the reasoning model of an expert. It is focused in bylaws, regulations, jurisprudence and customized background representing entity mission, vision and profile. This Structure and model are generic enough to self-adapt to any regulatory environment, but as a first step, it was limited to academic field. This way it is possible to limit the slang and data number. The foundations of the linguistic model are also outlined and the way the architecture implements the key features of the behavior.*

Keywords—Computational Linguistic, Linguistic Reasoning, Natural Language Processing, Text Mining, Chatter bot, Legal Advice, Semantics, Data Mining, Expert Systems

I. INTRODUCTION

During early ’50s, Alan Turing proposed the famous Turing test, one of the main challenges in the Artificial Intelligence field. That test intends to demonstrate the intelligence provided to a computer and, at the same time, the possibility that machines can think in a similar way humans do. [1]

J. Weizenbaum continued that idea but from other perspective, when builds a new soft engine afterwards named Eliza proyect[2]. Eliza is not just a computer program, but one of the first prototypes of early's Natural Language Processing (NLP). It implements a simple pattern matching as main strategy for understanding language, but constitutes one of the first conversational robots (chatter bots or chat bots). Some years later Dr. Colby creates Parry [6], a chat bot that mimics the behavior of a psychiatric patient suffering paranoia. It emulates responses according to different types of paranoia. Tests showed that at least a reduced set of psychiatrists were not able to distinguish between the computer and the human patient. [3]

Based on ELIZA [2], Richard Wallace developed a new project called Alice (1995) [5]. As one of the offsprings of this project, he also created AIML (Artificial Intelligence Mark-up Language), an application of XML, having a general label named Category, that constitutes the elemental unit of knowledge. Every category of knowledge has two components: Patterns and Templates. The pattern is a string of characters representing a dialog, and the template represents the answer to the pattern that is being activated. [7]

PTAH project has a chatter bot as part of the interface to the out-world, but also functions as a smart filter since its filters slang related to regulations and any legal instrument within academic scope.

It is important to find the proper context of the queries to be able to overcome problems like ambiguity, polysemy, anaphora etc. Most of the current solutions are based on approaches known as Natural Language Processing (NLP). [4] [8]

The solution typically involves one or more of the following levels of analysis: phonologic, morphologic, syntax, semantics, and pragmatics, but proposals rarely cover all of them at the same time. This layered approach is useful to break down the problem and make it simpler. Most of the times, there are larger dictionaries with certain per-processing that may be expensive or complex. Usually they become a corpus. They require certain human interaction in higher or less degree. [9]

Regarding the semantic frameworks (SFW), there are also many proposals that complement the previous initiatives. Among others WebODE [10][11][12], and ontological engineering, that allow to develop web sites to manage certain type of knowledge mostly automatically. Other SFW is ContentWeb, a platform for ontology integration with WebODE, that allows the user to interact using natural language but limited to certain slang[13]. That environment interacts with OntoTag (implemented with RDF/S and XML) [14], OntoConsult (interface for natural language based on ontology) and OntoAdvice (an information retrieval system based on ontology). Each word receives an URI (Uniform Resource

Identifier). It also defines an URI to every new morphosyntactic element.

There are environments to manage only morphosyntactics. For instance XTAG develops a pretty good English grammar [15]. It is based on a lexicalization model named Tree Adjoining Grammar (TAG), that generates a grammatical tree to solve the syntax processing. It includes a parser, an interface X-windows and a morphology analyzer.

As a last example, there is a tool for morphology [16] that performs morphological and syntactic analysis with disambiguated segmentation (splits text into segments according its coherence), special symbols disambiguation (used for sounds not related to words) and error correction for words misunderstood.

This work extends the chatter bot problem coordinating it with a linguistic model for reasoning in Spanish that leads the text processing from other perspective beside grammars clues. Therefore it introduces to the chatter bot, some extra technology related to Expert Systems, and semantic distance in a reduced field [23]. That technology was probed successfully in other contexts automatically managing context in a natural way [24]; specifically it was tested for automatic processing of Spanish dialogs [25].

Here, there is a discretionary morphosyntactic usage but is pending profiling and feature extraction from historical data using data driven approaches. This way the framework may be thought as a layered linguistic reasoning with two steps:

1. The first step is to filter sentences using a linguistic model. Classify the topic automatically using only clues (not semantic explicit by tags, nor structures or dictionaries). Those clues are morphosyntactic schemas previously learnt by an Expert System based on rules.

2. The second step allows a lightweight semantic association at word level using a non-metric distance implementing pre-defined relationships.

This article is organized in the following way: Section II presents the proposal to the problem of semantic distance related to similarity measurements; Section III shows a set of metric distances that may be used instead of the prototype option; Section IV presents the model PTAH, and conclusions and future work are shown in Section V.

II. Similarities Measured and Semantics

Searches using similarity have a large number of applications such as image and sound recognition, compression and text searching, computational biology, artificial intelligence and data mining among others [17]. All of them share the same characteristic: they look for similarities using certain distance or similarity function predefined for that case. The model most commonly used to represent data being processed is the metric space.

A metric space is defined as a couple (U, d) with U being the objects universe and d: U x U → R+ is a distance function defined for U elements that measure the similarity between them. That means the lower distance the closest the objects are. This function d follows also the typical properties for a metric distance:

$$\forall x, y \in U, d(x, y) \geq 0 \quad \text{(positive)} \quad \text{(eq.1)}$$

$$\forall x, y \in U, d(x, y) = d(y, x) \quad \text{(symmetry)} \quad \text{(eq.2)}$$

$$\forall x \in U, d(x, x) = 0 \quad \text{(reflexiveness)} \quad \text{(eq.3)}$$

$$\forall x, y, z \in U, d(x, y) \leq d(x, z) + d(z, y) \quad \text{(eq.4)}$$

(triangular inequality)

The database is a finite subset of the type X widely included in U with cardinality n. In this model, a typical query implies to retrieve similar objects using searches by certain ranks. Let them be d(q, r) with a query q ∈U and a tolerance radius r, a range search is to retrieve all the object s in the database that have a distance less than r from q. That is the same as in (eq. 5).

$$d(q, r) = \{x \in X / d(q, x) \leq r\} \quad \text{(eq.5)}$$

A searching by range can be solved with O(n) distance evaluations when examining exhaustively the Data Base (DB). To avoid that, it is possible to preprocess the DB by an algorithm that build an index to save time calculation at searching. An indexation algorithm is efficient if it can answer a query using similarity and making a minimum number of distance calculations, typically sub-lineal over the number of elements in the DB[18][19][20][21]. This project intends to query contents using similarity clues that improve semantic distance and require a lightweight algorithm.

III. Metrics for Distances

The previous metric analysis serves as an introduction on how a good distance must behave. Taking that into consideration, it is important to note that distances also strongly depend from the number and quality of the features that make part of the distances function. Among the most famous distances are the Euclidean and Manhattan [17]. But there are many others that are under consideration and evaluation as part of this project. They are depicted below just to show the scope of the global project. This part of the research is intended for flexibility evaluation of the model and also to make it clear how it can be improved. Some of the evaluated metrics were: Overlap Metric (OM), Value Difference Metric (VDM), Metric SFM and Minimum Risk Metric (MRM) [22].

IV. The Model and PTAH

The model implemented in PTAH has the modules shown in fig. 1.

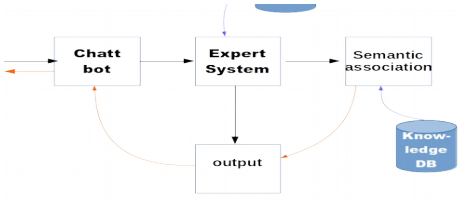

Fig. 1. Global architecture of PTAH. There is a connection between the Chatter Bot, the Expert System, the Semantic module and the DB.

a. Chatter Bot: it is the input module; a conversational robot coded in Python with patterns in AIML files. It responds to common conversations. From the input sentence, it selects the significant words of the query and removes the "stop words" obtaining a word set.

b. Expert System: implemented in Python too as a set of modules that derive the cases and topics. If the word set match a case, it submit the data to the Semantic Association module to search the documents in the Knowledge DB. The ES has a set of rules that outline the use cases of interest. A short list of them is in table II.

Table II. Detail of use cases and its relation to the ES rules

ID case	Use case	Rule
11.	¿Qué diferencia hay entre ser docente interino o concursado (ordinario)? What is the difference being transient teacher or regular teacher?	{cuándo |condiciones | diferencia} + {docente| profesor| auxiliar}+ {interino |concursado | ordinario}
13.	¿Cuál son las composiciones de los Consejos Departamental, Directivo y Superior? (How are composed the Department, Directive and Superior Councils?)	Consejo+{Departamental| Directivo| Superior}+ {composición |compuesto | miembros}

These rules are defined syntactically by the following CFG:

<Rule> ::= <AllList>
<Rule> ::= "{" <ExistList> "}"
<Rule> ::= "word"
<ExistList> ::= <ExistList> "|" <Rule>
<ExistList> ::= <Rule>
<AllList> ::= <AllList> "+" <Rule>
<AllList> ::= <Rule>

Then, rules are composed by words joined by "+" or "|" and grouped by "{" and "}". AllList are lists where all the components must be part of the query. ExistList represents components that at least one of them must be part of the word set. Words are compared by similarity using Levenshtein Edit Distance. If the distance is less than a radio, there is a match. We built an interpreter than check each rule against the word set and return the summation of the minimum distances that match the rule. A word set can match 0 or more rules ordered by distance.

If there is no match, the chatter bot use their AIML files to respond. Otherwise, the expert system sends the matching rules to the semantic association module.

c. Semantic Association Module: implemented as a set of stored procedures in PostgreSQL, its goal is to looks for the documents that have semantic similarity to the words set of the rule. Each rule has an optimized array indexed by word keys that represent the existence or not of the word key in the rule. An each document has a similar array with the frequency of the word keys in that document.

According to the model, the reasoning is represented by the following algorithm (semantic distance function):

1. Find Meta data(data)
2 .Find Binary vector(Meta data)
3. Retrieve binary vector using:
3. Bias A: restriction for no divergence
3.1. Binary vector from every bylaw
3.2. Relative frequency for every word (w) in the bylaw → freq(w)
3.3. MIN(w)= argMIN{ freq(w)}
3.4. MAX(w)= argMIN{ freq(w)}
3.5. find weighting for every w: p(w)=1/freq(w)
*3.6. p(w)=1/(MAX(w)*1.05)*
3. Bias B: relevance in the current case
3.1. Binary vector(ID-case)
3.2. NUM(w)= number of words in ID-case
3.3. Let $p'(w_i) = 1/NUM(w_i)$ for every w_i
3. Bias C: relevance in the current context
3.1. Binary vector(query(data))
3.2. For every w_i(Meta data) and every ID-case:
3.2.1. IF w_i(Meta data) AND w_i(ID-case) AND match (ID-case) scoring(ID-case)+= p'(w)
*3.2.2. IF w_i(Meta data) AND w_i(ID-case) ~match(ID-case) scoring(ID-case) += p'(w) * 0.95*
3.2.2.ELSE scoring(ID-case) - =p'(w) / there is no w_i(ID-case) in knowledgeDB */*
3.3.select argMAX{scoring(ID-case)}
3.3.1. IF number-of(ID-case) >1 THEN ID-caseBEST=select argMAX{freqH(ID-case)}
3. Bias D: hit precision in the DB
3.1. Search KnowledgeDB (binary vector(ID-caseBEST))
3.2. IF(hit (ID-caseBEST)) -> scoring += p(w)
*3.3. IF(~hit(ID-caseBEST) -> scoring += p(w) * 0.95*
3.4. ELSE / there is no w_i(ID-case) in knowledgeDB */*
3.4.1. IF hit(w_i(ID-case))-> scoring - = p (w)
3.4.2. IF ~hit (w_i(ID-case))-> scoring - = p' (w)
*3.5. Output (select * from KnowledgeDB where argMAX {scoring})*

In the algorithm, freqH represents the previous usage frequency, compiled during all the model history.

d. Knowledge DB: it is implemented over a PostgreSQL Data Base Management System. It is composed by a Documents table and an Articles table. One document can have 0 or more articles. Each article has an array of frequencies of word keys associated.

The DB is populated with textual information of the regulation, but it is expected to improve the data loading using also an OCR to include also non textual documents.

We already performed the first batch of experiments to determine the Precision and Recall of the system. Due to the restrictions about the extension of this short communication, we do not include these preliminary results.

V. CONCLUSIONS AND FUTURE WORK

This paper presents a linguistic reasoning for dialogs, compatible with the indirect semantic approach presented by

models using morphosyntactic but augmented with data driven-heuristics. The PTAH prototype implements that model extending the traditional processing for chatter bot using new layers of abstraction that do not fit in the traditional strategies of NLP.

Those layers distribute filters among an ES rule-based and the following explicit steps:

Bias A: restriction for no divergence

Bias B: relevance in the current case

Bias C: relevance in the current context

Bias D: hit precision

It is important to note that it is not required the labeling, dictionaries or trained corpus. From preliminary results, it can be seen that PRECISION and RECALL metrics are pretty good even though the distance metric is poor and can be improved with better distance functions.

As pending tasks it can be mentioned the following:

- Add dictionaries and historical data to improve query results
- Self tuning of the rules in the ES. Also the rules could be learnt probabilistically from history.
- Evaluate other metric distances that may evidence linguistic relationships between words. It would improve newer situations and make the system more flexible
- Evaluate the precision and recall with higher number of queries
- Evaluate the same parameters using the distances in section III.
- Implement a new module for OCR and automatic loading of the DB.
- Improve the interface using a synthesizer and a voice recognition system. This could make the interaction more friendly.
- Extend the use cases to other topics improving the chatter bot to be less sensible to slang and dialects.
- Enhance the ruled system with Fuzzy Logic.

REFERENCES

[1] Turing, Alan, “Computing Machinery and Intelligence”, Mind 59: 433-60, (1950).

[2] J. Weizenbaum. “ELIZA- A Computer Program for the Study of Natural Language Communication between Man and Machine”, Communications of Association for Computing Machinery 9, 36-45, (1966):.

[3] J.Weizenbaum. “Computer power and human reason”. San Francisco, CA. W.H. Freeman, (1976).

[4] T. Winograd. “Procedures as a Representation for Data in a Computer Program for Understanding Natural Language”. Cognitive Psychology Vol.3 Nº 1, (1972).

[5] ALICEBOT, alicebot.blogspot.com/

[6] K.M Colby, F.D. Hilf, S. Weber, J. Kraemer. "Turing-Like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes". A.I., 3, 199-222, (1972).

[7] R.Wallace. “The Elements of AIML Style”. ALICE AI FOUNDATION, (2003).

[8] Manning C., Schütze H. “Foundations of Statistical Natural Language Processing”. MIT Press. (1999).

[9] M. Mauldin. “Chatterbots, TinyMuds and The Turing Test: Entering The Loebner Prize Competition”. AAAI-94, (1994).

[10] Corcho O., López Cima A., Gómez Pérez A. “A Platform for the Development of Semantic Web Portals”. ICWE’06. Palo Alto. California. USA. ACM. (2006).

[11] Corcho O., Fernández-López M., Gómez-Pérez A. Vicente O. “WebODE: an integrated workbench for ontology representation, reasoning and exchange”. Knowledge Engineering and Knowledge Management Proceedings. Volume: 2473, 138-153, (2002).

[12] http://webode.dia.fi.upm.es/WebODEWeb/index.html

[13] Aguado de Cea G., Álvarez de Mon y Rego I., Pareja Lora A. “Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: Onto Tag". Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. No. 17. pp. 37 - 49. (2002).

[14] Aguado de Cea G., Álvarez de Mon y Rego I., Pareja Lora A., Plaza ArtecheR."RFD(S)/XML LINGUISTIC ANNOTATION OF SEMANTIC WEB PAGES". International Conference on Computational Linguistics. Proceedings of the 2nd workshop on NLP and XML - Volume 17.pp 1 - 8. (2002).

[15] Paroubek P., Schabes Y., Joshi A. K. “XTAG - A Graphical Workbench for Developing Tree-Adjoining Grammars”. Third Conference on Applied Natural Language Processing, Trento (Italy), (1992).

[16] Prószéky G., Naszódi M., Kis B. “Recognition Assistance: Treating Errors in Texts Acquired from Various Recognition Processes”. International Conference on Computational Linguistics. Proceedings of the 19th international conference on Computational linguistics - Volume 2.pp 1 – 5, (2002).

[17] E. Chavez, G. Navarro, R. Baeza-Yates, and J.L. Marroquın. Searching in metric spaces. ACM Computing Surveys, 33(3):273–321, September (2001).

[18] R. Baeza-Yates, W. Cunto, U. Manber, and S. Wu. Proximity matching using fixed-queries trees. In Proc. 5th Combinatorial Pattern Matching (CPM’94), LNCS 807, pages 198–212, (1994).

[19] Chavez, E., Marroquin, J., and Navarro, G. 2001a. Fixed queries array: A fast and economical data structure for proximity searching. Multimed. Tools Appl. 14, 2 (June), 113–135. (Expanded version of Overcoming the curse of dimensionality. In European Workshop on Content-Based Multimedia Indexing, 57–64, Toulouse, France, (1999).

[20] Bugnion, E., Fel, S., Roos, T., Widmayer, P., and Widmer, F. A spatial index for approximate multiple string matching. In Proceedings of the 1st South American Workshop on String Processing (WSP’93) (Belo Horizonte, Brazil), R. Baeza-Yates and N. Ziviani, Eds. 43–53. (1993)

[21] Chavez, E. and Navarro, G. An effective clustering algorithm to index high dimensional spaces. In Proceedings String Precessing and Information Retrieval (SPIRE 2000) (A. Coruna, Spain), 75-86. (2000)

[22] Chaoqun Li and Hongwei Li. A Survey of Distance Metrics for Nominal. Attributes. JOURNAL OF SOFTWARE, VOL. 5, NO. 11. Department of Mathematics, China University of Geo-sciences, Wuhan, Hubei, China 430074. Email: fchqli, hwlig@cug.edu.cn, (2010).

[23] López De Luise, D. “Morphosyntactic Linguistic Wavelets for Knowledge Management“. Libro "Intelligent Systems", ISBN 979-953-307-593-7. InTech OPEN BOOK. (2011).

[24] López De Luise, D, Hisgenr, D. “Language Modeling with Morphosyntactic Linguistic Wavelets ”. . "Automatic content extraction on the web with intelligent algorithms". CIDM. (2013).

[25] López De Luise, D. Hisgen, A. Cabrer, M. Morales Rins."Modeling dialogs with Linguistic Wavelets". IADIS Theory and Practice in Modern Computing 2012 (TPMC 2012). Lisboa, Portugal. (2012).