Natural Language Understanding: from laboratory predictions to real interactions

Pedro Mota, Luísa Coheur, Sérgio Curto, Pedro Fialho
L^{2}F / INESC-ID Lisboa
Lisboa Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
{pedro.mota, luisa.coheur, sergio.curto,
pedro.fialho}@l2f.inesc-id.pt

Abstract. In this paper we target Natural Language Understanding in the context of Conversational Agents that answer questions about their topics of expertise, and have in their knowledge base question/answer pairs, limiting the understanding problem to the task of finding the question in the knowledge base that will trigger the most appropriate answer to a given (new) question. We implement such an agent and different state of the art techniques are tested, covering several paradigms, and moving from lab experiments to tests with real users. First, we test the implemented techniques in a corpus built by the agent's developers, corresponding to the expected questions; then we test the same techniques in a corpus representing interactions between the agent and real users. Interestingly, results show that the best "lab" techniques are not necessarily the best for real scenarios, even if only in-domain questions are considered.

Key words: Natural language understanding, classification, information retrieval, string similarities, laboratory experiments, real interactions

1 Introduction

Natural Language Understanding (NLU) targets at mapping given utterances into some sort of representation that models their meaning and that a computer is able to process. Intense research has been dedicated to NLU, as it is in the basis of the development of many applications. Conversational Agents – agents that interact in natural language – are examples of such applications.

In this paper, we study the NLU process of question-answering agents, that is, agents that target to teach, ask or answer questions about a topic in which they are "experts" [2, 7, 14]. In addition, we focus on the particular case of the agents whose knowledge base is constituted by sets of pre-defined interactions [9, 11, 12]. In all these agents, the NLU task consists in, being given a question, finding the "closest" question in the knowledge base, so that its answer is appropriate to the new question. Although only appropriate in limited domains, this approach has the advantage of allowing the fast development of a prototype, even by non experts, as the knowledge sources are simple question/answers pairs [12].

In this paper, different state of the art techniques – representing what has been applied to such agents – are compared, both in a lab environment and also in a real scenario of interaction with the agent. Interestingly, results show that the best "lab" techniques are not necessarily the best for real scenarios, even if only in-domain questions are considered.

The paper is organised as follows: in Section 2 we present related work, in Section 3 we describe the surveyed NLU techniques, in Section 4 we report and discuss the results of our experiments. Finally, in Section 5 we draw some conclusions and point to future work.

2 Related Work

NLU aims at mapping utterances into some sort of representation that models their meaning and that the system knows how to interpret. The mapping must correctly capture the meaning of the given utterance and, thus, must be able to map to the same representation, utterances with the same meaning. There are many possible approaches to this task [4, 15, 16].

NLU is in the basis of the performance of many applications, including Conversational Agents. Although some of these are task oriented, others target to give information about a specific domain. Examples of such virtual agents are Max [6], Hans Christian Andersen (HCA) [1], Sergeant Blackwell (SB), Edgar [12] and DuarteDigital [11]. Agent Max is deployed at the Heinz Nixdorf MuseumsForum, in Germany, and its main goal is to provide explanations and comments on the museum's exhibits. HCA personifies the renown fairy tale author in a computer game and is able to talk about his work. SB plays the role of a military instructor that is prepared to talk about the US Army and answers the questions of potential recruits [7]. Both DuArte Digital and Edgar answer questions about art, in Portuguese: Edgar's topic of expertise is the Monserrate Palace and DuArte Digital's is a piece of jewelry from the 16th century, Custódia de Belém.

While Max and HCA have sophisticated NLU modules that output, respectively, conversational acts and different levels of categories (syntactic, semantic, etc.), SB, Edgar and DuArte Digital have in their knowledge sources pairs constituted by utterances (typically questions) and the respective answers. It should be clear that all utterances associated with the same answer should have the same meaning or, at least, should be related somehow so that each utterance from that set can be answered by the same sentence. Therefore, within these agents, the NLU task is to be able to map a given sentence into some "paraphrase" existing in their knowledge bases.

SB's treats NLU as a *Text Classification* problem [9], as studied in the Information Retrieval (IR) field, with a few differences. In IR the set of possible answers is perceived as a collection of documents and the strategy used consists in viewing the query and a document as samples from some probability distribution over words and make a comparison between those distributions. This approach is not suitable for SB because it is common to have question/answer pairs in which there are no common words (the vocabulary mismatch problem). Motivated by this, a new approach was developed by Leuski and his team. This approach is called Cross Language Model (CLM) and corresponds to a statistical approach in which the vocabulary mismatch problem is overcome by assuming that there are two distinct languages, one for the questions and another for the answers. Therefore, the answers need to be "translated" into the question language, or the other way around, before comparing the word distributions. A detailed description on how CLM works can be found in [9]. This same approach is also applied to these agents [10, 13].

Considering Edgar's NLU capabilities, they are modelled as a classification process, which is a relatively common approach for NLU (see, for instance, [3]). As described by Edgar's authors, and considering that its knowledge base has utterances/answers pairs in the form ($\{u_{i1}$, ..., $u_{in}\}$, a_i), where each u_{ij} represents a possible way of formulating an utterance to which a_i is an appropriate answer, the NLU process can be seen as a classification task, if we assume that each pair of the knowledge base represents a distinct category. That is, all the utterances u_{i1}, ..., u_{in} correspond to the same category.

Finally, DuArte Digital is "modeled as a service which, for each user question, searches for the most similar question".

3 Implemented Natural Language Understanding Techniques

Being given a set of utterances and respective answers about a specific topic representing the existing knowledge base and a new (unseen) utterance, we target to identify the utterance in the knowledge base that is "closer" to the new utterance, so that we can appropriately answer to it. This corresponds to the classification task described in agent Edgar (Section 2). An example of this NLU process is to have in the knowledge base the entry *($\{$What is your name?, Can you tell me your name?$\}$, My name is Edgar)* and associate the input *Have you been given a name?* to it. In the end the corresponding answer is returned. As stated before, in the literature this task can be performed using different paradigms, such as Information Retrieval classification or as a lexical similarity operation. In the following we detail the methods which were used in this paper.

In the basis of our work is the platform described in [12], which is publicly available. This platform allows us to treat NLU as a classification process (as defined in Section 2), and uses a Support Vector Machines (SVM) to this end, by using the `LIBSVM` [5] library, with the most appropriate kernels and optimized parameters, in a one-versus-all multi-class strategy.

We have extended this platform by allowing the expansion of the training data with synonyms or paraphrases, which are defined separately. For instance, if the utterance $w_1...w_i...w_j...w_n$ exists in the corpus and is stated that $w_i...w_j$ is the same as $s_k...s_l$, then the utterance $w_1...s_k...s_l...w_n$ is added.

Also, we add to this platform the possibility of dealing with NLU as done in CLM and a method based on string similarity measures.

Considering the CLM approach, SB's developers have created the NPCEditor toolkit (NPC stands for Non Player Character) [8], which implements the CLM. Thus, the virtual human kit[1] containing NPCEditor was used, through the provided message API.

[1] http://vhtoolkit.ict.usc.edu/index.php

In what respects string similarity, we opted for three measures with distinct natures: Jaccard, Overlap and Dice. Measures such as Levenshtein were not considered as sentences with the same tokens occurring in a different order are strongly sanctioned by these.

Jaccard (Equation (1)) obtains higher scores for utterances that have similar length (a zero value means that there is nothing in common between two sentences; one is the highest possible value).

$$Jaccard(U_1, U_2) = \frac{|U_1 \cap U_2|}{|U_1 \cup U_2|} \tag{1}$$

A different philosophy is used in the Overlap measure (Equation (2)), in which sentences with different lengths are not so strongly penalised. A combination of the Jaccard and Overlap measures (Equation (3)) was implemented. In this strategy an empirical weighting factor λ is used.

$$Overlap(U_1, U_2) = \frac{|U_1 \cap U_2|}{min(|U_1|, |U_2|)} \tag{2}$$

$$\begin{aligned} JaccardOverlap(U_1, U_2) =& \lambda \times Jaccard(U_1, U_2) \\ &+ (1 - \lambda) \times Overlap(U1, U2) \end{aligned} \tag{3}$$

Following the perspective of trying to balance the length of the sentences, we also chose the Dice similarity measure (Equation (4)) that does not favour any particular size of utterance.

$$Dice(U_1, U_2) = 2 \times \frac{|U_1 \cap U_2|}{|U_1| + |U_2|} \tag{4}$$

In order to use either of the previously described string similarity measures for the necessary NLU task, we compute the similarity score between the input utterance and the utterances in knowledge base entries. In the end, the answer of the utterance in which the highest score was obtained is returned. Also, any n-gram version of the utterances can be used in the similarity measures.

4 Experiments

In this section we report the results of our experiments where we have tested the different approaches previously described. These experiments were made in the context of the virtual agent Edgar.

4.1 Experimental setup

Three distinct corpora were used in the experiments (details of these corpora can be seen in Table 1):

- LAB_UNEXPANDED: built by two experts in the topic of our agent;

- LAB_EXPANDED: corresponds to an expansion of LAB_UNEXPANDED. This expansion was made using a set of 154 rules;
- REAL: gathered through a user test scenario, via web. It was told to the participants which was our agent's topic of expertise; then users were instructed to freely interact with the system. We had 6 participants in this evaluation, 3 male and 3 female, from 25 to 40 years old. No post-processing of the gathered input was made.

	LAB_UNEXPANDED	**LAB_EXPANDED**	**REAL**
Categories	124	124	26
Number of utterances	603	6181	112
Max utterances in a category	20	1080	-
Min utterances in a category	1	1	-
Number of words	3976	46456	611
Unique words	519	618	186

Table 1. Details of the different corpora.

In what concerns the REAL corpus, there are 56 and 60 out of vocabulary words, concerning, respectively, the LAB_UNEXPANDED and the LAB_EXPANDED corpora. After collecting this corpus, a manual classification of the utterances was made, in order to divide individual interactions into three possible categories: `in-domain`, `out-of-domain` and `context`. Utterances are considered in-domain if they relate to a topic to which the agent is prepared to answer; they are out-of-domain, otherwise, except if they fall into the `context` category, which occurs if the utterance requires contextual information from previous interactions in order to be answered. Table 2 summarises this process.

	In-domain	**Out-of-domain**	**Context**
Number of utterances	72	34	6
Number of words	390	209	12
Unique words	121	102	9

Table 2. Characteristics of the REAL corpus.

4.2 Testing with "Lab" corpora

First experiments were done with LAB_UNEXPANDED and LAB_EXPANDED corpora. A 5-fold cross validation procedure was followed, that is, the corpora was divided in 5 random partitions and NLU techniques were trained with 4 partitions and tested with the remaining one. At the end, an average of the results was made. These results can be seen in Table 3. As expected, higher accuracy scores are obtained for all NLU techniques in the LAB_EXPANDED corpus. A reason for this is the low number of utterances in some categories in the LAB_UNEXPANDED corpus; another reason derives from the fact that many of the utterances with the same meaning were lexically very distance from each other; worse, many utterances with different meanings (and thus, quite different) were grouped together, since they could all be answered by the same utterance. Therefore, when the division of training and test corpus was made, we ended up with utterances in the test set that were completely different from the ones in the training, leading to poorer results in the LAB_UNEXPANDED corpus.

NLU technique	LAB_UNEXPANDED	LAB_EXPANDED
SVM binary unigrams	0.71 ± 0.040	0.98 ± 0.004
CLM	0.73 ± 0.049	0.97 ± 0.008
Overlap bigrams	0.66 ± 0.051	0.96 ± 0.021
Jaccard unigrams	0.67 ± 0.021	0.97 ± 0.006
Jaccard Overlap unigrams	0.69 ± 0.034	0.97 ± 0.006
Dice bigrams	0.69 ± 0.034	0.97 ± 0.005

Table 3. Accuracy results of the cross validation procedure on the agent Edgar's corpus.

The almost perfect results in the LAB_EXPANDED corpus can be explained by the fact that there are many utterances corresponding to the same category. Thus, in this scenario, utterances in the test set are very likely to have a similar utterance in the training set.

4.3 Testing with a "Real" corpus

For this experiment we only considered the in-domain questions of the REAL corpus. NLU techniques were trained with the LAB_UNEXPANDED and LAB_EXPANDED corpora. Here, we wanted to evaluate how the techniques under study performed in a user test scenario. Results can be seen in Table 4.

Two interesting facts that can be observed:

- In a real user scenario the NLU techniques that performed better were the string similarity ones. This was not expected because in the lab experiments these were not the best techniques;
- Results for the LAB_EXPANDED corpus were still better (a 9% increase in the best case), but the difference of performance between the techniques decreased.

4.4 Discussion

From a more detailed analysis of results, we highlight the following points:

NLU technique	LAB_UNEXPANDED	LAB_EXPANDED
SVM unigrams	0.60	0.69
CLM	0.65	0.68
Overlap bigrams	0.56	0.66
Jaccard unigrams	0.64	0.72
Jaccard Overlap unigrams	0.67	0.72
Dice unigrams	0.64	0.71

Table 4. Accuracy results for in-domain questions made in the user evaluation.

- CLM is influenced by the number of utterances per meaning. That is, if a set U_1 of utterances with meaning M_1 share some words with another set of utterances U_2 with meaning M_2 and if the number of utterances in U_1 is (much) larger than the number of utterances in U_2, then every time these words appear, M_1 is much more likely to be chosen;
- The SVM is not sensible to the previous situation, but due to its discriminative nature, it can be biased if some word/expression only occurs in a set of utterances with the same meaning (category). In this situation, if a given utterance contains this word/expression it will most probably be mapped into that meaning;
- String similarity measures have a different problem: they just allow comparisons between utterances, two by two, and the whole picture is lost.

To conclude, lab experiments, in the case of the LAB_UNEXPANDED corpus, serve as a predictor of the NLU techniques performance in a real scenario, as the discrepancies between accuracy results of both experiments are not very high. Thus, although these observations might be biased by the low number of users, we believe that these experiments can be useful in estimating the results of NLU techniques in a user evaluation scenario.

5 Conclusions and Future Work

In this paper we tested several NLU techniques, which are applied by conversational agents that have in their knowledge base utterances/answer pairs. Two main experiments were done: one involving a lab scenario, corresponding to evaluating the system with questions selected by its developers; in the second experiment, techniques were evaluated with real users. The analysis of both experiments revealed that the best technique in lab is not necessarily the best in a real scenario and that it is possible to roughly estimate how NLU techniques will behave in a user evaluation. Drawbacks of the different techniques were also identified.

Concerning future work, since each technique has specific characteristics and different reasons for succeeding and failing, we will study ways of combining them. Dealing with out-of-domain questions is another important issue, because it is something inevitable in a real scenario. Finally, we will continue to make improvements in our prototype in order to move to other NLU approaches.

Acknowledgments

This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through the project FALACOMIGO (ProjectoVII em co-promoção, QREN n 13449) that supports Pedro Mota, Sérgio Mendes and Pedro Fialho's fellowships.

References

1. Niels Ole Bernsen and Laila Dybkjær: Meet hans christian anderson, In: *Proceedings of the Sixth SIGdial Workshop on Discourse and Dialogue*, pp. 237–241, (2005).
2. Niels Ole Bernsen, Marcela Charfuelan, Andrea Corradini, Laila Dybkjær, Thomas Hansen, Svend Kiilerich, Mykola Kolodnytsky, Dmytro Kupkin, and Manish Mehta: Conversational H.C. Andersen first prototype description, In: *ADS*, pp. 305–308, (2004).
3. Rahul Bhagat, A. Leuski, and Eduard Hovy: Shallow semantic parsing despite little training data, In: *Proc. ACL/SIGPARSE 9th Int. Workshop on Parsing Technologies*, (2005).
4. Dan Bohus and Alexander I Rudnicky:Ravenclaw : Dialog management using hierarchical task decomposition and an expectation agenda, In: *Phoenix Usa*, 4–7, (2003).
5. Chih-Chung Chang and Chih-Jen Lin, In: *LIBSVM: a library for support vector machines*, 2001.
6. Lars Gesellensetter, Nicole C. Kramer, and Ipke Wachsmuth: A conversational agent as museum guide - design and evaluation of a real-world application, In: *The 5th International Working Conference on Intelligent Virtual Agents*, pp. 329–343. Springer, (2005).
7. Anton Leuski, Jarrell Pair, David R. Traum, Peter J. McNerney, Panayiotis P. Georgiou, and Ronakkumar Patel: How to talk to a hologram, In: *IUI*, eds., Cécile Paris and Candace L. Sidner, pp. 360–362. ACM, (2006).
8. Anton Leuski and David Traum: NPCEditor: A Tool for Building Question-Answering Characters, (2010).
9. Anton Leuski and David R. Traum: A statistical approach for text processing in virtual humans, In: *Proceedings of the 26th Army Science Conference*, (2008).
10. Anton Leuski and David R. Traum: Npceditor: Creating virtual human dialogue using information retrieval techniques, In: *AI Magazine*, pp. 242–256, (2011).
11. Ana Mendes, Rui Prada, and Lusa Coheur: Adapting a virtual agent to users vocabulary and needs, In: *Intelligent Virtual Agents*, eds., Zsfia Ruttkay, Michael Kipp, Anton Nijholt, and Hannes Vilhjlmsson, volume 5773 of *Lecture Notes in Computer Science*, 529–530, Springer Berlin / Heidelberg, (2009).
12. Catarina Moreira, Ana Cristina Mendes, Luísa Coheur, and Bruno Martins: Towards the rapid development of a natural language understanding module', In: *IVA*, pp. 309–315, (2011).
13. Thomas D. Parsons: Affect-sensitive virtual standardized patient interface system, In *Clinical Technologies: Concepts, Methodologies, Tools and Applications*, volume 3, (2011).
14. Thies Pfeiffer, Christian Liguda, and Ipke Wachsmuth: Living with a virtual agent: Seven years with an embodied conversational agent at the Heinz Nixdorf MuseumsForum, In: *kk*, 273–297, (1995).
15. A. Rudnicky and Xu W.: An agenda-based dialog management architecture for spoken language systems, In: *IEEE ASRU Workshop*, pp. 337–340, (1999).
16. Joseph Weizenbaum, Eliza - a computer program for the study of natural language communication between man and machine, In: *Communications of the ACM*, **9**(1), 36–45, (1966).