Natural Language Processing in IBM Watson Assistant, an Automatic Verification Process

[1,2]Beatriz Consciência, [1,2]Simão Santos, [1]Pedro Martins, [2]Luís Pombo, [1]Steven Abrantes, [1]Cristina Wanzeller

[1]Polytechnic Institute of Viseu, Dep. of Computer Sciences, Viseu, Portugal
{pedromom, steven, cwanzeller}@estgv.ipv.pt
[2]Sofinsa, IBM, Portugal
{beatriz.consciencia, simao.santos, jose.luis.pombo}@pt.softinsa.com

***Abstract*—The exponential growth of Artificial Intelligence powered systems affect us all. IBM Watson Assistant is one of the central AI-powered systems used in small and larger business. For instance, in Portugal, IBM Watson is powering call-center systems for companies in banking and telecommunications business. This paper proposes a strategy that can automate the verification process of generalization capability in the creation of chat-bots using IBM's platform. K-Fold cross-validation is a favorite technique in machine learning for estimating the performance of a learned hypothesis on a data set. Therefore, the proposed method is not new for testing. However, this method is newly applied to the chat-bot application using IBM's platform. In this paper, the primary goal is to make the chat-bot testing process automated, with the objective of making it faster, more productive, and efficient. Algorithms like k-fold cross-validation demonstrate the need for a representative and reasonable amount of data when it comes to training IBM Watson in his ability to learn.**

Keywords-IBM Watson; cognitive computing; neural networks; natural language processing.

I. Introduction

As said, the progress in artificial intelligence is incredibly fast, close to exponential [13]. Companies are adopting this technology so they can make the most of it. For instance, call centers, hotel's room services, car's dashboards or even information balconies in airports are being redesigned to embrace AI.

In 2010, IBM presented Watson, a supercomputer whose architecture was question-answer based [15]. Recently, in March 2018, IBM launched IBM Watson Assistant, a platform that allows every user to create his very own chat-bot within the domain of knowledge that the user intends the chat-bot to be fluent at. Using this platform, chat-bots are built and deployed in a few clicks in a very user-friendly graphical interface. For that matter, users only need to acknowledge few terms and concepts such as "intent", "entities", "dialogue" and "digressions". In the upcoming section, those concepts are described.

With IBM Watson Assistant, users can deploy their chat-bots on other platforms such as Slack or even Facebook. However, when building a chat-bot application for an enterprise, it needs to be tested before going into production phase and, that testing should be carried out by real-life users who interact with it and point out the failures on recognition and misunderstandings of user intentions. The main goal is to make chat-bot testing an automated process so that it becomes a lot faster and consequently more productive and efficient for the business.

This research work presents a solution for the automation of the testing process in the creation of chat-bots in the Portuguese language, with the IBM Watson Assistant platform.

This paper is organized as follows. Section II, makes a brief resume of the related work. Section III, describes basic background knowledge regarding chat-bots. Section IV explains, making use of some examples how the k-fold algorithm is applied. Section V, identifies the case study used to motivate this work. Section VI, briefly describes, work-in-progress, implementation. Finally, section VII, point out some of the conclusions and future work directions.

II. Related work

IBM Watson Assistant was presented on March 19, 2018, at the "IBM Think 2018" in Las Vegas. It is a branch of the giant IBM Watson aimed at creating assistants intuitively [6].

Virtual assistants created with the help of Watson Assistant are targeted for business environments, which means that its main ambition is not to be available directly to the general public. [4]. Watson Assistant has already been tested in the new Maserati GranCabrio dashboard computer and at the Munich airport, integrated with a robot, destined to answer to transient issues [2].

Besides, the IBM Watson Assistant, as stated earlier, is possible to integrate into any and every business [3]. For example, in a hotel, this can be integrated into each room and allow the customer through simple commands to control the temperature and ambient lights or even ask for room services [5]. Another example, in the case of a bank, makes it possible to integrate IBM Watson Assistant as a chat-bot on the institution where the customer can carry out operations without the need of replacing the concept of "customer support" [10]. However, IBM Watson Assistant, as well as IBM Watson, do not communicate in Portuguese of Portugal. There is not yet an algorithm of natural language comprehension oriented to Portuguese, given the complexity and ambiguity of the language.

 ISBN: 978-1-61208-670-5

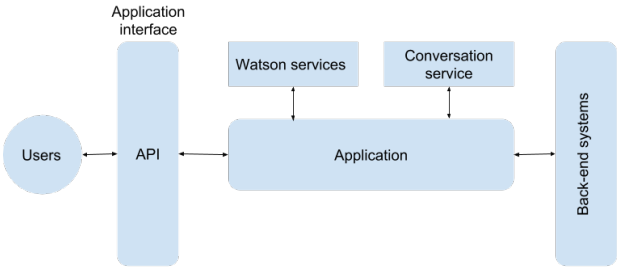

Fig. 1: IBM Watson Assistant integration

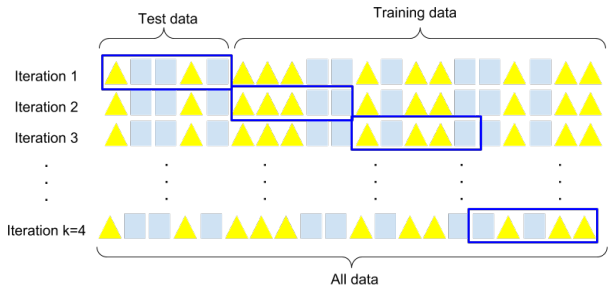

Fig. 2: Samples Partition into k equal sized sub-samples. Image adapted from [14].

Watson is also an assistant based on an intent architecture [16], where intents represent the user's intention in his or her speech, that is, what he or she desires and can usually be identified by a verb like "to go" or "to do" [12].

To the best of our knowledge, in the field of Portuguese (Portugal) chat-bots, this area is uncharted.

The Portuguese language, today is the sixth most spoken language in the world, with 244 million speakers spread across five continents. Thus, there is a large market for the Portuguese language that has not been yet conquered by AI-powered chat-bots.

Watson Assistant proposes that its users will be able to, in a short time, create a chat-bot, with which they can establish a dialogue. That creation is done in three steps, corpus creation, design, and integration [7]. Figure 1 shows a scheme of IBM Watson Assistant integration.

III. Concepts of chat-bots architecture

As said before, some concepts need to be understood before diving into chat-bot development. Most of those concepts are regarding "intents" and "entities." Intents represent the intention of a user when he/she addresses a chat-bot. For instance, if someone says "I want to buy a book", wanting to buy was the intention and therefore it would trigger an intent that contains other phrases related to that intention, and that can also trigger it. Another key term concerning to chat-bots is "entity". Referring again to the example sentence, "I want to buy a book", the entity, in this case, would be "books". So, entities are words that put intents into perspective giving them a purpose.

The other two terms that are relevant when it comes to developing a chat-bot with the IBM Watson Assistant platform are "dialogue" and "digressions". The first term refers to the dialogue flow which the chat-bot need to follow to have a structured speech. The second term, "digressions", refers to the conditions which allow the user off the chat-bot to jump from one intent to another and return to the former one if intended. Those represent a set of rules that allow the chat-bot to change context without losing the sense of the speech.

IV. K-fold Cross Validation

K-fold Cross Validation is algorithm mainly used in machine learning related problems where the main goal is to predict something and/or to estimate the reliability of a predictive model [1]. It is mostly applied to Supervised Learning models where a data-set is given with known labels of the target, on which training is performed (training data-set) [9]. It is also given a data-set of unlabeled data, against which the model is tested (testing set) [8]. The purpose of cross-validation is to test the model's ability to predict new data that were not used in estimating it, to flag problems like over-fitting and to give an insight on how the model will generalize when deployed on production [17].

To test and validate the proposed system, the original sample is randomly partitioned into k equal sized sub-samples, as can be seen in Figure 2. Of the k sub-samples, a single sub-sample is retained as the validation data for testing the model, and the remaining $k - 1$ sub-samples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k sub-samples used exactly once as the validation data. The k number of results from the folds can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each representation is used for validation precisely once [11].

Given the example "I want to buy a book", that was previously classified as related to specific intent, the k-fold algorithm will match that previous classification with the classification resulting from IBM Watson Assistant. After that, it will present the result of that comparison (equal or not equal). This approach is then applied to the entire data model (i.e., the test array). When all instances have been compared, the k-fold algorithm will provide the results for each metric (e.g., Accuracy) per k fold which can be used to generate synthetic parameters such as the average accuracy of the whole chat-bot.

V. Case Study

For this study, a chat-bot was developed in IBM Watson Assistant platform. That chat-bot was a banking related one, therefore named Credit-bot.

Credit-bot is a chat-bot created just for this study on an academic proposal. It allows performing simple tasks that usually require a trip to the local bank branch. For instance, making a bank wire transfer; open an account, or even making a credit simulation. Notice, again, this tasks were simulated for academic purposes only.

 ISBN: 978-1-61208-670-5

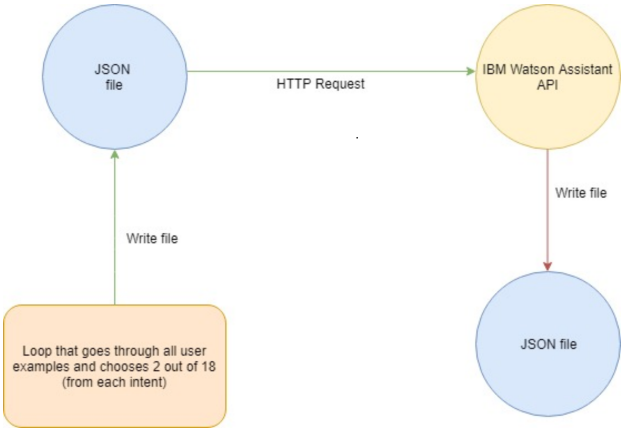

Fig. 3: Implementation process

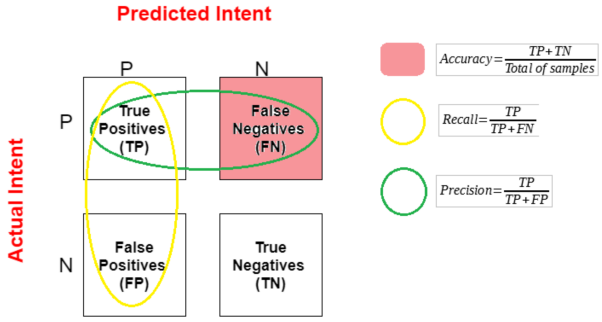

Fig. 4: Metrics

To make Credit-bot learn information about this domain some intents where created such as “Making Transfer” that compiles a set of useful user examples. User examples are phrases the user may address to the chat-bot and which need to relate to the respective intent to follow a workflow. Eleven intents to perform banking related actions were created, and twenty user examples were given to each intent, making a total of 220 user examples. Besides that, five entities were also created containing female names, male names, Portuguese city’s, villages, terms related to account types and credit types.

This case-study it is aimed to create a chat-bot able of answering in Portuguese (Portugal) natural language to most frequent users requests. For instance, when using a chat, the user can authenticate himself, and then request in natural language (written) to transfer some money from his account to any other account. The same way, he can request credit, loans, pay regular bills, or even create routine tasks.

VI. Implementation

After the chat-bot creation in IBM Watson Assistant platform, it was produced a JSON file with the user examples needed to feed the chat-bot. This JSON file was then sent to the IBM Watson Assistant platform that reads the file and trains the chat-bot. For this study, the 20 examples per intent were hand-crafted (in a total of 220 examples). For each intent, the similar examples were split, where 18 were assigned to training data and 2 for testing. The process of sending this arrays to IBM platform is also performed through a JSON file. The process was then repeated ten times, as ten is the number of k. All samples were then sent to the IBM Watson Assistant through the API, where they were classified. This process scheme can be seen in Figure 3.

As a result of the classification, the API outputs a classification JSON file that was used by the k-fold algorithm. That JSON file was the target of an examination by the algorithm itself to compare the classification made by the IBM Watson Assistant versus the expected rating. That comparison resulted in another classification in another JSON file being the parameters “True Positive” when the classification made by the system was equal to the expected, and “False Positive” when it was not. Moreover, when an example was classified with “True Positive” the other examples must be classified as “True Negative”, and when an example was classified as “False Positive” the other must be classified as “False Negative”.

That classification was then used to calculate parameters like accuracy, recall, precision, and F1 score. This classification and obtained results are based on a confusion matrix. The method used to reach the metrics are depicted in Figure 4. The experimental results can be seen in Table I.

The IBM Watson Assistant platform was then powered with a set of test arrays. Each test array consisted on a set of examples, frequently named “user examples”, that are phrases that a user might address to IBM Watson Assistant to trigger an intent. Each test array had in total 22 user examples. The test arrays were then used to test the platform generalization ability when classifying examples that were unknown. This test was performed to analyze where the model fails, so that model gaps are recognized and corrected. Gaps include, for example, spelling errors, unstructured sentences or even slang words, are translated into intents (e.g., if the user addresses to IBM Watson the sentence “send money” the system must recognize it with the intent “Make a transfer”) that require a greater quantity or quality of data for disambiguation.

Results in Table I, show that it is possible for the proposed model to identify almost all true positives (TP). Thus, having a high Recall value (close to 1). However, there is also a high number of false positives (FP), which tells us that in many cases the platform classification fails, assuming the test cases as being relative to an intent of which they are not.

As for the means of precision by intent, these tell us that two intents have more miss-classification by the platform, such as attempting to “AffirmativeResponse” (positive answer) and “MakeTransfer” (make transfer). These intents are therefore more susceptible to ambiguity than the rest, which may lead to a dialogue with many doubts, making it unpleasant to the user. One possible reason for the ambiguity is the similarities in the training-set.

As far as the averages by k are concerned, from the same statistical distribution, presented values are equal, independently of the data subset, allowing no other conclusions than the one obtained.

 ISBN: 978-1-61208-670-5

TABLE I: RESULTS FROM METRICS

INTENT	AVERAGEPRECISION	STDPRECISION	AVERAGEACCURACY	STDACCURACY	AVERAGERECALL	STDRECALL	AVERAGEF1SCORE	STDF1SCORE
AFIRMATIVERESPONSE	0.4	0	0.945	0	1	0	0.571	0
MAKETRANSFER	0.4	0	0.945	0	1	0	0.571	0
FAREWELL	0.417	0.029	0.947	0.003	1	0	0.588	0.028
CAPACITIES	0.438	0.048	0.949	0.004	1	0	0.608	0.046
BOT_CONTROL_CHANGE_SUBJECT	0.45	0.05	0.95	0.005	1	0	0.619	0.048
BOT_CONTROL_CLARIFICATION	0.458	0.049	0.951	0.004	1	0	0.627	0.047
CREDITSIMULATION	0.464	0.048	0.951	0.004	1	0	0.633	0.045
ACCOUNTBALANCE	0.469	0.046	0.952	0.004	1	0	0.637	0.044
NEGATIVERESPONSE	0.472	0.044	0.952	0.004	1	0	0.64	0.042
OPENACCOUNT	0.475	0.042	0.952	0.004	1	0	0.643	0.04
GREETING	0.477	0.041	0.952	0.004	1	0	0.645	0.039

Main conclusions point out that “AffirmativeResponse” and “Make a transfer” has a lower generalization capability when compared to other intents. As stated above, this is because the given examples are very similar.

VII. Conclusions

The main goal of this paper was to outline a strategy to test the quality of a chat-bot automatically. This was achieved through means of the application of the k-fold cross-validation algorithm on the outputs of Watson Assistant intent classification, which allow us to assess the generalization ability of the underlying model through multiple metric which, in turn, give us a strong indication on the quality of the chat-bot solution. In practice, the advantages include the ability to identify in which topics the chat-bot will have more trouble figuring out the user true intent (therefore correcting the situation by adding more relevant data or splitting intents) and a way to continuously improve the quality of the chat-bot by continuously providing real-world usage data while measuring the quality improvements of the model. For instance, the chat-bot had less generalization ability towards the intent “MakeTransfer”. This can be concluded through the values of average accuracy, that for that intent are ”0”, meaning that there was no record of true positive neither true negatives which means that there has the only record of false negatives and false positives. There is, none user example trigger the intent correctly. This is related to the quality of data given to the chat-bot since example phrases related to the entities do not variate much in content. Therefore we must find other ways to invoke the intent and consequently feed the intent with more examples so it can generalize better. This also happens with the intent “AffirmativeResponse”.

Future work will include semantic analysis, integration of grammar parsing tools, text and/to speech recognition and other business tools provided by IBM Watson.

Other future work directions, lead to apply the proposed system to help visually impaired people, as well as, hearing impaired people.

Acknowledgements

“This article is financed by national funds through FCT - Fundação para a Ciência e Tecnologia, I.P., under the project UID/Multi/04016/2016. Furthermore, we would like to thank the Instituto Politécnico de Viseu for their support.”

More, we especially would like to thank the Viseu Polytechnic professors: Prof. Cristina Wanzeller for her outstanding work supervision; Prof. Pedro Martins, for his support and orientation. Finally, we thank Softinsa, IBM group, for making possible this research work.

References

[1] M. W. Browne and R. Cudeck. Single sample cross-validation indices for covariance structures. *Multivariate behavioral research*, 24(4):445–455, 1989.

[2] H. Chen, R. H. Chiang, and V. C. Storey. Business intelligence and analytics: from big data to big impact. *MIS quarterly*, pages 1165–1188, 2012.

[3] Y. Chen, J. E. Argentinis, and G. Weber. Ibm watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clinical therapeutics*, 38(4):688–701, 2016.

[4] L. Comerford, D. Frank, P. Gopalakrishnan, R. Gopinath, and J. Sedivy. The ibm personal speech assistant. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 1–4. IEEE, 2001.

[5] F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, and S. Weerawarana. Unraveling the web services web: an introduction to soap, wsdl, and uddi. *IEEE Internet computing*, 6(2):86–93, 2002.

[6] R. High. The era of cognitive systems: An inside look at ibm watson and how it works. *IBM Corporation, Redbooks*, 2012.

[7] P. Imrie and P. Bednar. Virtual personal assistant. In *ItAIS 2013*. AIS Electronic Library (AISeL), 2013.

[8] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen. Gesture-based affective computing on motion capture data. In *International conference on affective computing and intelligent interaction*, pages 1–7. Springer, 2005.

[9] R. Kohavi. Glossary of terms. *Special issue on applications of machine learning and the knowledge discovery process*, 30(271):127–132, 1998.

[10] A. P. Massey, M. M. Montoya-Weiss, and K. Holcom. Re-engineering the customer relationship: leveraging knowledge assets at ibm. *Decision Support Systems*, 32(2):155–170, 2001.

[11] G. McLachlan, K.-A. Do, and C. Ambroise. *Analyzing microarray gene expression data*, volume 422. John Wiley & Sons, 2005.

[12] S. Memeti and S. Pllana. Papa: A parallel programming assistant powered by ibm watson cognitive computing technology. *Journal of Computational Science*, 2018.

[13] M. Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.

[14] F. Pedregosa-Izquierdo. *Feature extraction and supervised learning on fMRI: from practice to theory*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2015.

[15] C. Thompson. What is ibm's watson. *New York Times Magazine (accessed on October 2018). http://www. nytimes. com/2010/06/20/magazine/20Computer-t. html*, 2018.

[16] C. Todd, R. Vazquez Pena, and R. Srinivas. Evaluation of artificial intelligence frameworks. *SMU Data Science Review*, 1(1):10, 2018.

[17] X. Xie, J. W. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, 84(4):544–558, 2011.

 ISBN: 978-1-61208-670-5