Performance Comparison of Natural Language Understanding Engines in the Educational Domain

Víctor Juan Jimenez Flores[1]
Faculty of Engineering
Universidad José Carlos Mariátegui
Moquegua, Perú

Oscar Juan Jimenez Flores[2]
Faculty of Engineering
Universidad Privada de Tacna
Tacna, Perú

Juan Carlos Jimenez Flores[3]
Contracts and Services
Southern Peru Copper Corporation
Tacna, Perú

Juan Ubaldo Jimenez Castilla[4]
Faculty of Engineering
Universidad José Carlos Mariátegui
Moquegua, Perú

***Abstract*—Recently, chatbots are having a great importance in different domains and are becoming more and more common in customer service. One possible cause is the wide variety of platforms that offer the natural language understanding as a service, for which no programming skills are required. Then, the problem is related to which platform to use to develop a chatbot in the educational domain. Therefore, the main objective of this paper is to compare the main natural language understanding (NLU) engines and determine which could perform better in the educational domain. In this way, researchers can make more justified decisions about which NLU engine to use to develop an educational chatbot. Besides, in this study, six NLU platforms were compared and performance was measured with the F1 score. Training data and input messages were extracted from Mariateguino Bot, which was the chatbot of the José Carlos Mariátegui University during 2018. The results of this comparison indicates that Watson Assistant has the best performance, with an average F1 score of 0.82, which means that it is able to answer correctly in most cases. Finally, other factors can condition the choice of a natural language understanding engine, so that ultimately the choice is left to the user.**

***Keywords*—Chatbot; natural language understanding; NLU; F1 score; performance**

I. Introduction

Nowadays, the business and researchers are progressively perceive the importance of chatbot systems, because they are integrated into daily life, playing roles as assistants to end users [1]. In the educational domain, Kowalski [2] indicates that chatbots can play an important role, because it represents an interactive mechanism, instead of the traditional e-learning systems, where students can constantly ask questions related to a specific field.

On the other hand, most research does not emphasize the used natural language understanding (NLU) engine, or its choice is not very justified. Therefore, this research compares different NLU engines, like Google Dialogflow, Microsoft LUIS, IBM Watson Assistant, Wit.ai, Amazon LEX and Rasa (an open source chatbot framework) and tries to answer, in terms of performance and educational domain, which one to use.

A chatbot is a computer program which uses machine learning technique voice recognition, and natural language processing (NLP) to conduct a intelligent conversation with a person (e.g. Amazon's Alexa and Google's Assistant) [3]. Moreover, one of the main components of a chatbot is the natural language understanding (NLU), which is the ability of a machine to understand human languages, said otherwise, it is the process of converting natural language text into a form that computers can understand [4].

In addition, multiple researches and comparisons were made regarding the different natural language understanding engines available, as mentioned in Section II; however, their limitation was that they used test data. This research tries to fill that gap.

In this research, training data and input messages were extracted from Mariateguino Bot,which was the chatbot of the José Carlos Mariátegui University during 2018, its main function was to attend to the doubts of the students regarding the necessary requirements to carry out administrative procedures. Moreover, the chatbot was able to answer frequently asked questions, support students regarding the admissions process, and provide class schedules.

Mariateguino Bot was made with Dialogflow, a Google service that runs on Google Cloud Platform. Therefore, other platforms, described in Section III, were evaluated.

To determine the performance of the evaluated services, the F1 score was used. F1 score is a performance measure for compare the quality of predictions between systems [5]. In the same way, F1 score is defined as the harmonic mean of precision and recall [6]. It was chosen because it is one of the most practical ways to numerically calculate the performance of an NLU engine and it is widely used in related researches. In addition, according to [7], using the f1 score, the results can be easily compared with previous works, because it is one of the standard metrics for performance measurement.

Finally, this paper is divided into seven sections. Section II gives a brief overview of related works. Section III defines the NLU engines evaluated during the research. Section IV describes the methodology. Section V and section IV describe

the results and discussions. Finally, Section VII and Section VIII describe the conclusions and future work.

II. Related Works

In recent years, many researches have been carried out regarding chatbots and the impact they have on traditional processes, generally in customer service. Some performance related researches are listed below.

Canonico and De Russis wrote a paper titled "A comparison and Critique of Natural Language Understanding Tools" [8], which compares the main cloud-based platforms, from a descriptive and performance based point of view. Their results showed that Watsson Assistant is the platform who performs best.

On the other hand, Braun, Hernandez, Matthes and Langen wrote a paper titles "Evaluating Natural Language Understanding Services for Conversational Question Answering Systems", which presents a method to evaluate the classification performance of NLU services. Their results indicated that LUIS showed the best scores and RASA could achieve similar results.

Unlike the aforementioned researches, this paper makes a comparison between six natural language understanding engines, with input messages and training data data that belonged to a real chatbot from the José Carlos Mariátegui University. Also, the main language of the chatbot was Spanish.

III. Natural Language Understanding Engines

There are many natural language understanding modules that are available as cloud services and major IT players like Google, Microsoft, IBM, Facebook and Amazon have created tools to develop chatbots [9]. Additionally, Rasa was included because Dialogflow training data can be converted to its format and is an open source alternative compared to the other platforms.

A. Dialogflow

According to Sabharwal and Agrawal [10], Dialogflow is one of the services from Google Compute Platform that makes it easy to integrate cognitive virtual agents to traditional applications; also, it uses natural language understanding and natural language processing capabilities to build complex uses cases.

B. LUIS

The Language Understanding Intelligent Service (LUIS) is a Microsoft's bot engine that runs on Azure Cognitive Services [11].

C. Watsson Assistant

The Watson Assistant service enables learning to respond to the customers in a way that simulates a conversation between humans [12]. In addition, Watsson Assistant is a IBM's bot engine.

D. Wit.ai

Wit.ai is a Facebook's bot engine which allows training bots with sample conversations and have your bots repeatedly learn from interrelating with customers [13].

E. Amazon LEX

Amazon Lex is a Amazon's bot engine for building intelligent assistants or chatbots, which provides many AI capabilities like Automatic Speech Recognition (ASR) and Natural language Understanding (NLU) [14].

F. Rasa

Rasa NLU is an open-source NLP library for intent classification and entity extraction in chatbots [15].

IV. Methods

The method of evaluating the classification performance of NLU engines is based on [16].

A. Materials

The NLU engines evaluated during the research were Dialogflow, Wit.ai, LUIS, Amazon LEX and Rasa. Moreover, 100 messages from the Mariateguino Bot conversation history were randomly selected as input data and they were grouped based on the expected intents, thus obtaining 30 intents. To calculate the performance of each platform, the F1 score metric was used, which includes precision and recall. Some input messages from the students and the expected intents are shown in Table I.

TABLE I. Input message examples

Expected Intent	Input Message
req-tramites	solicito constancia de no adeudo
req-tramites	Record académico oficina
INF-universidad-telefonos	Cual es el numero de telefono de servicios academicos de moquegua
INF-universidad-telefonos	Teléfono de UJCM moquegua
INF-universidad-pago-mensualidades	Fecha de pago de pensiones
INF-universidad-pago-mensualidades	Costo de matrícula
INF-universidad-matriculas	hasta cuando me puedo matricular
INF-universidad-inicio-clases	¿Cuando inician las clases?
FAQ-lugar-clases	¿En dónde serán las clases de Derecho?
FAQ-lugar-clases	Donde sera las clases de ingenieria comercial
FAQ-docentes	Docentes de este ciclo?
FAQ-docentes	Plana docente
FAQ-docentes	Profesores de educación
Default Welcome Intent	Hola
Default Fallback Intent	Administracion VI ciclo
Default Fallback Intent	Quiero estudiar de cero
Default Fallback Intent	Quiero saber si hay carreras a distancia

```
{
  - luis: {
        message: "¿Cuando inician las clases?",
        intent: "Default Fallback Intent",
        confidence: 1
    },
  - watson: {
        message: "¿Cuando inician las clases?",
        intent: "inf-universidad-inicio-clases",
        confidence: 0.8172303199768067
    },
  - witai: {
        message: "¿Cuando inician las clases?",
        intent: "inf-universidad-inicio-clases",
        confidence: 0.6489
    },
  - lex: {
        message: "¿Cuando inician las clases?",
        intent: "faqlugarclases",
        confidence: "?"
    },
  - dialogflow: {
        message: "¿Cuando inician las clases?",
        intent: "INF-universidad-inicio-clases",
        confidence: 0.8454428911209106
    },
  - rasa: {
        message: "¿Cuando inician las clases?",
        intent: "inf-universidad-inicio-clases",
        confidence: 0.8966025710105896
    }
}
```


Fig. 1. Node.js Application Output

B. Procedure

As a first step, a conversion of the Dialogflow training data to the rest of the research platforms was carried out, using the QBox.ai service, available in https://qbox.a. Then, one hundred messages were randomly selected from Mariateguino Bot conversation history and they were grouped based on the expected intents, thus obtaining 30 intents.

Afterwards, to start testing and obtain data, a Node.js application was created in order to combine the application programming interface (API) from each NLU engine, in such a way that each input message was only entered once and the desired data was obtained in the format shown in Fig. 1. Also, a threshold of 0.5 was programmed for all platforms, so that if the API of the NLU engine returns a confidence less than 0.5, the Node.js application returns the default fallback intent.

In order to evaluate the results, the predicted intent were identified for each input message. In this way, true positives (TP), false positives (FP) and false negatives (FN) were calculated.

As a final step, the performance of the NLU engines was measure in terms of precision, recall and F1 score, given by the following expressions:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1 score is defined as the harmonic mean of precision and recall [6].

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

These measures were applied for single intents, then the average F1 score was calculated. For this research, one NLU engine is better than another if it has a higher average F1 score.

V. Results

The results shown in Fig. 2, Fig. 3, Fig. 4 and Table II are the average precision, recall and F1 score of the 30 intents that were evaluated for each natural language understanding engine.

In terms of precision, as Fig. 2 shows, Dialogflow has the highest value (0.83), while LUIS obtained the lowest value (0.46). This means that the majority of cases that Dialogflow marked as positive, were correct.

In terms of recall, as Fig. 2 shows, Watson Assistant has the highest value (0.89). This means that Watson Assistant correctly identified the majority of positive cases from the total number of cases. On the other hand, LUIS obtained the lowest value (0.34).

Finally, in terms of F1 score, calculated from precision and recall, as Fig. 2 shows, Watson Assistant and Dialogflow have the highest value (0.82), while LUIS obtained the lowest value (0.36). The possible cause of the low performance of Microsoft LUIS is discussed in the section VI.

Overall, as can be seen in Table II, Watson Assistant and Dialogflow performed better, while LUIS obtained the lowest performance.

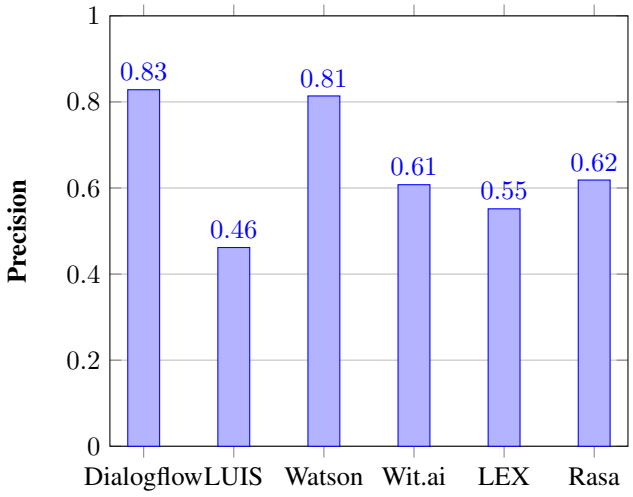

Fig. 2. Precision of Natural Language Understanding Engines

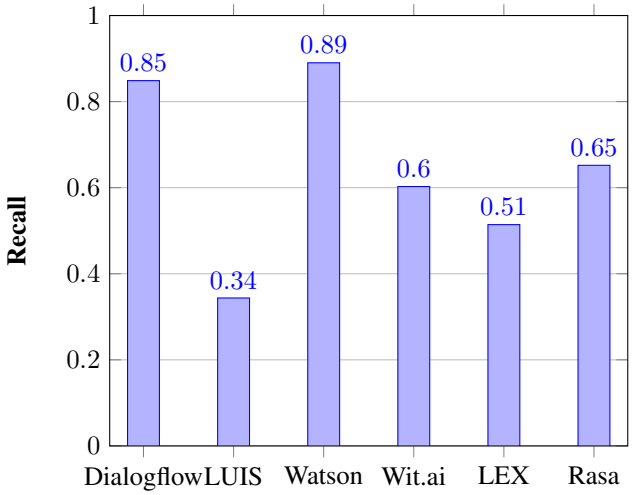

Fig. 3. Recall of Natural Language Understanding Engines

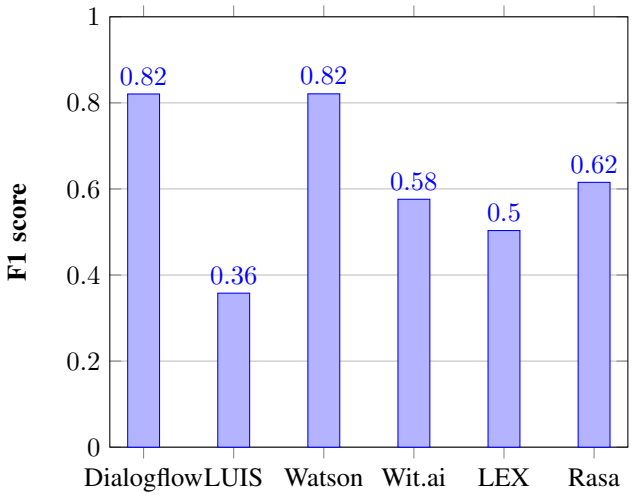

Fig. 4. F1 Score of Natural Language Understanding Engines

TABLE II. F1 SCORES OVERVIEW

NLU Engine	Precision	Recall	F1 score
Dialogflow	0.83	0.85	0.82
LUIS	0.46	0.34	0.35
Watson Assistant	0.81	0.89	0.82
Wit.ai	0.61	0.60	0.58
Amazon LEX	0.55	0.51	0.50
Rasa	0.62	0.65	0.62

VI. Discussion

Despite the fact that Watson Assistant and Dialogflow obtained the same F1 score, Watson Assistant can be considered performed best because the original service with which the chatbot was in production was Dialogflow, so it was constantly improving only on that service.

On the other hand, the low performance of LUIS may be due to the language. Mariateguino Bot was a chatbot made for students in the Spanish language and, despite the fact that LUIS has Spanish in its configuration, it was observed that the intent classification decreases considerably in the presence of input messages that have words with a Spanish accent.

Lastly, the final goal of this research was to compare the main natural language understanding engines and determine which one has the highest performance in the educational domain. Watson assistant was the service with the highest performance; however, for [16], LUIS showed the best results. This difference may be due to the fact that the chatbot domain and language was not the same. Moreover, we agree that Rasa can get better results, after some customization, because, during the present research, its full potential as an open source solution was not exploited. In addition, we agree with [8], which indicates that Watson is the platform that performs best since it can assign the correct intention in most of the cases studied, with a high confidence level.

VII. Conclusion

This study presented a performance comparison of Dialogflow, LUIS, Watson Assistant, Wit.ai, Amazon LEX and Rasa services in the educational domain, in order to determine which chatbot solution performs best and provide future researchers with more information on which service to choose. It was concluded that Watson Assistant showed the best performance and its use is recommended for the development of chatbots belonging to the educational domain. However, other factors may affect the choice of a platform that provides the NLU engine, such as the level of usability of the service or pricing plans. Therefore, it will be the company or researcher who decides which service best suits their needs.

On the other hand, the performance obtained by Rasa can be considerably improved with the appropriate settings, keeping in mind that it is an open source chatbot framework with a powerful natural language understanding engine.

VIII. Future Work

As future work, we plan to evaluate the performance of NLU engines across multiple domains. Similarly, we plan to evaluate the optimal threshold in order to improve performance, since for this research, we only worked with 0.5.

Acknowledgment

We would like to thank the José Carlos Mariátegui University, for allowing us to put Mariateguino Bot into production, and its students, for their feedback.

REFERENCES

[1] J. J. Bird, A. Ekárt, and D. R. Faria, "Learning from interaction: An intelligent networked-based human-bot and bot-bot chatbot system," in *UK Workshop on Computational Intelligence*. Springer, 2018, pp. 179–190.

[2] S. Kowalski, R. Hoffmann, R. Jain, and M. Mumtaz, "Universities Services in the New Social Ecosystems: Using Conversational Agents to Help Teach Information Security Risk Analysis," in *SOTICS 2011, The First International Conference on Social Eco-Informatics*, 2011, pp. 91–94.

[3] A. Mittal, *Getting Started with Chatbots: Learn and create your own chatbot with deep understanding of Artificial Intelligence and Machine Learning*. Bpb Publications, 2019.

[4] N. Pathak, *Artificial Intelligence for .NET: Speech, Language, and Search: Building Smart Applications with Microsoft Cognitive Services APIs*. Apress, 2017.

[5] Z. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal Thresholding of Classifiers to Maximize F1 Measure," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II*. Springer, 2014, p. 715.

[6] O. Campesato, *Artificial Intelligence, Machine Learning, and Deep Learning*. Mercury Learning & Information, 2020.

[7] G. Arnicans, V. Arnicane, J. Borzovs, and L. Niedrite, *Databases and Information Systems: 12th International Baltic Conference, DB&IS 2016, Riga, Latvia, July 4-6, 2016, Proceedings*, ser. Communications in Computer and Information Science. Springer International Publishing, 2016.

[8] M. Canonico and L. De Russis, "A comparison and critique of natural language understanding tools," *Cloud Computing*, vol. 2018, p. 120, 2018.

[9] P. Hall, V. Venigalla, and S. Janarthanam, *Hands-On Chatbots and Conversational UI Development: Build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. Packt Publishing, 2017.

[10] B. Galitsky, *Developing enterprise chatbots : learning linguistic structures*. Springer, 2019.

[11] N. Pathak and A. Bhandari, *IoT, AI, and Blockchain for .NET: Building a Next-Generation Application from the Ground Up*. Apress, 2018.

[12] S. Vetter, A. Azraq, S. Chughtai, A. Mashhour, D. V. Nguyen, R. M. Dos Santos, and I. B. M. Redbooks, *Enhancing the IBM Power Systems Platform with IBM Watson Services*. IBM Redbooks, 2018.

[13] J. Seligman, *ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING AND MARKETING MANAGEMENT*, 2018.

[14] S. Tripuraneni and C. Song, *Hands-On Artificial Intelligence on Amazon Web Services: Decrease the time to market for AI and ML applications with the power of AWS*. Packt Publishing, 2019.

[15] S. Raj, *Building Chatbots with Python: Using Natural Language Processing and Machine Learning*. Apress, 2018.

[16] D. Braun, A. Hernandez, F. Matthes, and M. Langen, "Evaluating natural language understanding services for conversational question answering systems," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 174–185.