Platform Spotlight

Benchmarking the Chatlayer NLP in 2021

01. Introduction

As an Enterprise SaaS platform provider, it's incumbent on us to challenge ourselves and our technology on a regular basis. Continuous improvement and value-based roadmap development are parts of our DNA, so understanding how our in-house NLP (Natural Language Processing) Engine compares to major players in the AI field is very important to us.

As our proprietary NLP is at the heart of our Conversational AI platform, its quality impacts the entire solution, all the way down to how a customer feels when interacting with our chatbots. To ensure that our NLP continues to perform on or exceeding par, we've benchmarked it against IBM Watson, Microsoft LUIS and Google DialogFlow, based on an independent and relevant dataset, across multiple languages. The goal of this benchmark was to assess intent classification accuracy in English, Portuguese, Spanish, French, and Dutch.

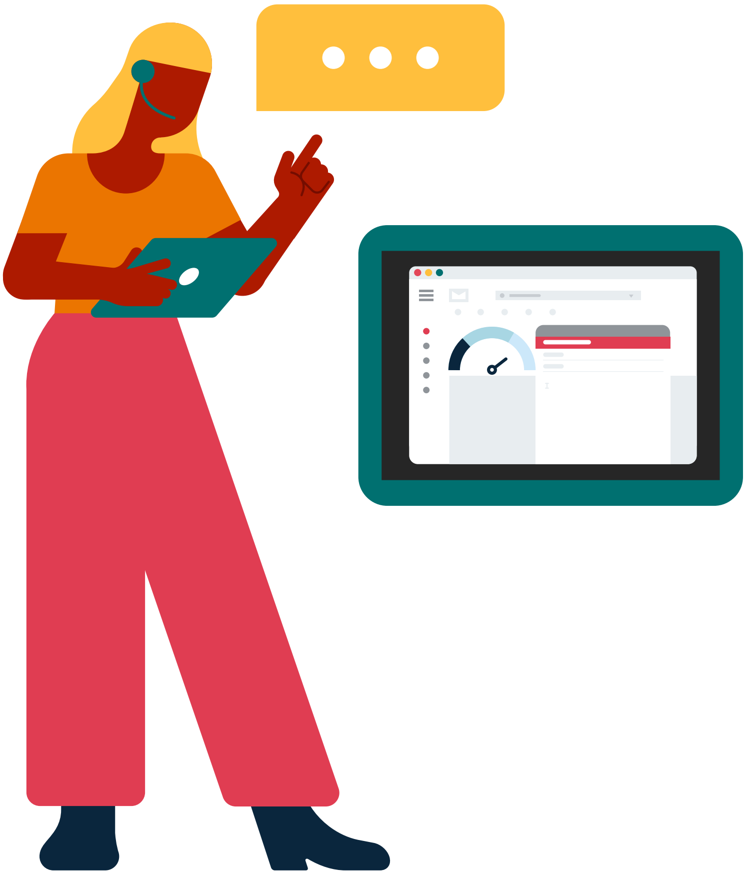
02. The benchmark

For testing, we used an independent dataset, Banking77. This dataset contains 77 intents used for creating a chatbot for the banking industry - a relevant domain for Chatlayer, considering the many banking and insurance clients using the Chatlayer platform. But, besides knowing how to effectively answer within topics it has been trained on, chatbots should be able to detect in-domain questions that are not part of the predetermined questions that it supports.

For example, a bot in the banking industry might be able to answer questions about transferring money, but not about using your credit card abroad. Additionally, it should also be able to detect and handle out-of-domain questions like “what is the color of the sky?”. Hence, to make the dataset more realistic, we filtered this set so that 10% of intents were removed from training to ensure negative examples (in-domain questions that cannot be answered) in the test set. We also added 5% out-of-domain questions.

When we query the bot with the abovementioned not supported in-domain questions and out-of-domain questions, the most desirable outcome is that it detects an intent with low confidence, which allows us to notice that the bot is not prepared to reply to that question.

Thus, if we properly tune the confidence threshold, the bot will return fallback responses in those cases (“sorry, I don’t understand”), which is better than having the bot return unwanted responses.



03. The results

The results of the English benchmark are profound, with Chatlayer (84%) outperforming Microsoft (79%), Google (82%) and IBM (81%) in intent classification (Table 1). Our main metric is Accuracy, which measures how many of all queries were properly handled by the bot.

Intent classification results for English

platform	accuracy	F1-Score
Chatlayer	84	86
Watson	81	84
Luis	79	81
Dialogflow	82	84

Table 1.

What about other languages?

We translated the above dataset into four other languages: Spanish, French, Dutch and Portuguese. When comparing each dataset across the languages tested, we can see that Chatlayer outperforms the three competing platforms in all five languages (Fig. 1).

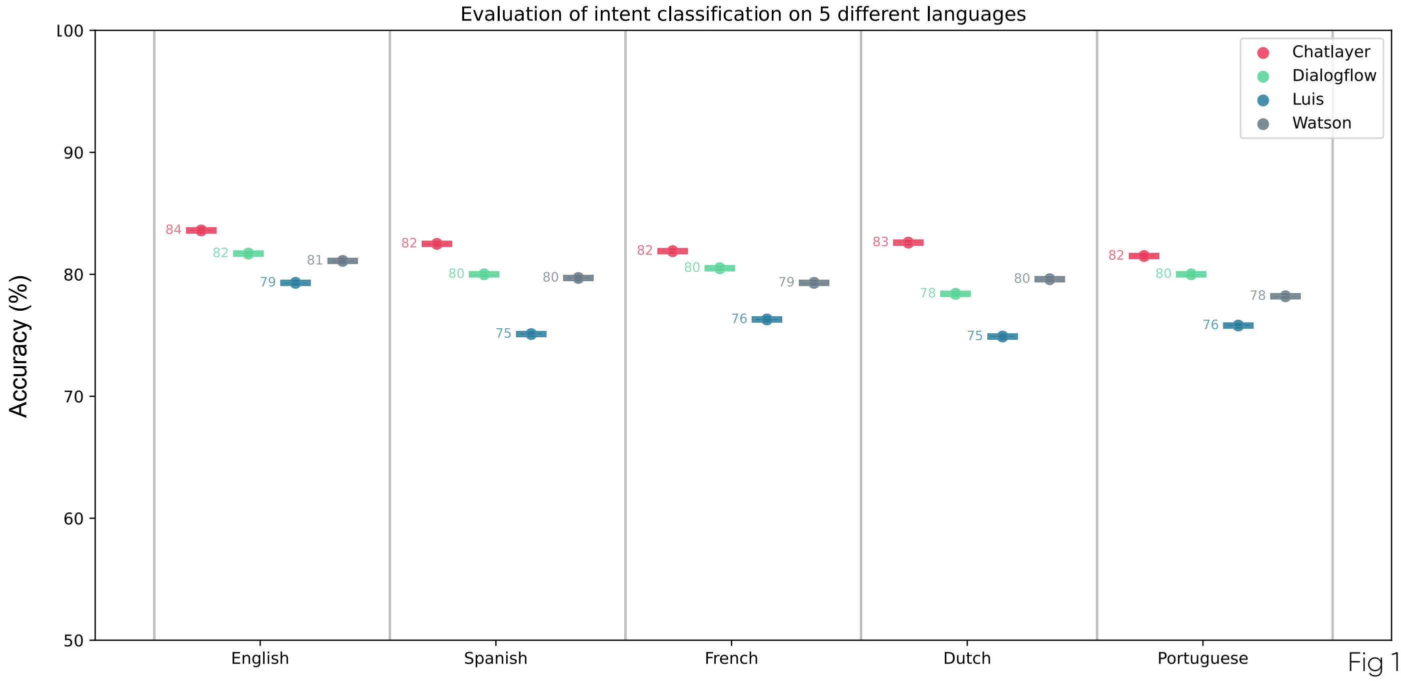


Fig 1.

What if you do not have a lot of data?

We wondered: what would the outcome be if you only have 10 training examples per intent, or 20, 30, 40 or 50? Which platform performs best? As you can see in Fig. 2, Chatlayer still outpaces the competition, with the lead being the biggest when you only have a few examples. This is great for companies that want to bootstrap chatbots with minimum effort.

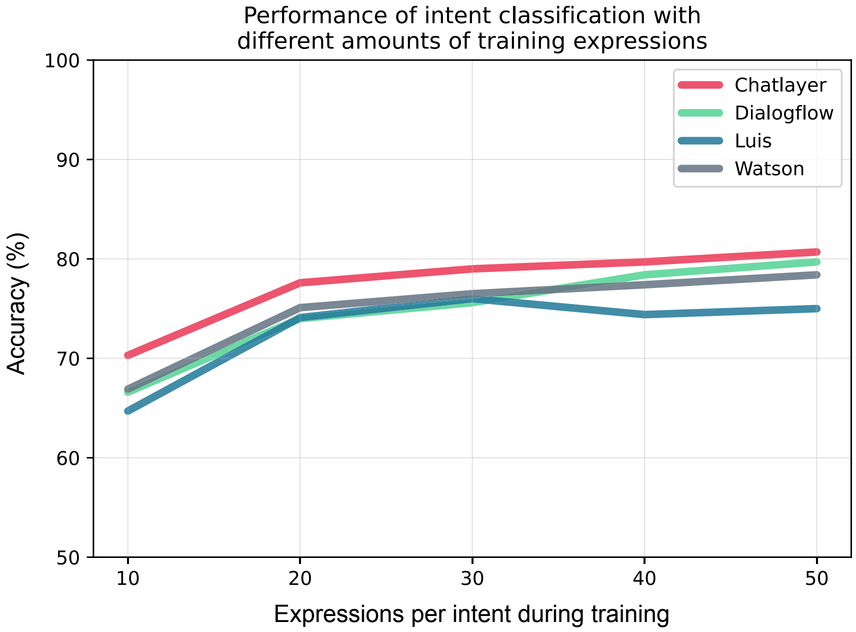


Fig. 2

Additional notes

While these results speak for themselves, our engineering teams additionally conducted real-world tests to validate Chatlayer Confidence and Intent classification, achieving similar results and solidifying our leading position in this benchmark. Also, these benchmarks were run on our current production engine as this provides accurate results for the present day – when running the benchmark on our pre-production V2 engine, slated for release in the fourth quarter of 2021, the accuracy was even higher across all languages. Once V2 enters production, we will reconduct this benchmark study and update our results, but in initial testing, the Chatlayer V2 engine achieves an accuracy of 86% and F1 of 89% on English with their latest Transformer models (see Table 1 for comparison).

04. Conclusion

When it comes to Intent classification accuracy, Chatlayer outshines leading AI platforms in real-world scenarios across multiple languages. And this is a lead we intend to keep. Through effective innovation, Chatlayer will continue to push its platform to the highest levels of performance, delivering increased quality and results to the end user. In 2022, our intent recognition will become fully context-aware resulting in even better NLU capabilities for chatbots.