bitext we help AI understand humans

Benchmark of Dialogflow and Bitext performance

Introduction

One of the main problems with the current generation of chatbots is that they require large amounts of training data. If you want your chatbot to recognize a specific intent, you need to provide it with a large number of sentences that express that intent. Until now, these large training corpora had to be generated manually. This is a time-consuming task rather than a creative one, and it makes the success of bot development very costly. To solve this problem, at Bitext we offer our Artificial Training Data service, which automatically generates many different sentences with the same meaning as the original one, in order to automate the most resource-intensive part of the bot creation process.

Dialogflow is one of the most popular chatbot-building platforms, for that reason we have chosen it for our tests. We tested how Dialogflow can benefit from the Artificial Training Data approach, comparing bots trained with hand-tagged sentences with bots extended with no effort with automatically-generated training data. Our tests show that if we train with only 2 or 3 example sentences per intent in Dialogflow, performance suffers. When we train with 10 sentences per intent, the results improve only slightly. In contrast, by extending these hand-tagged corpora with additional variants, automatically generated with Artificial Training Data, we get a significant improvement and higher accuracy overall.

Madrid, Spain
José Echegaray, 8 - Building 3, Office 4, 28232 Las Rozas

San Francisco, California
541 Jefferson Ave., Ste. 100 Redwood City, CA 94063

What we have done

We have carried out two different tests (A and B). Both use the 5 following intents related to the management of lights in a house:

- Switch on lights (switch on the lights in the living room)
- Switch off lights (switch off the lights in the living room)
- Change the color of lights (change the lights to blue)
- Dim lights (dim the living room lights to 20%)
- Program lights for a specific time (program the garden lights for 21:00).

In both tests, we have also used the same 5 types of slots: ACTION, OBJECT, PLACE, PERCENTAGE and HOUR.

In the first test (A), we trained two different bots. A first bot (A1) was trained with only 12 hand-tagged sentences (2 to 3 sentences per intent). Using those sentences as input, our Bitext Artificial Training Data service generated 391 sentences which, together with the 12 sentences from bot A1, were used to train a second bot A2 (with around 80 sentences per intent).

To evaluate both A1 and A2, we used an evaluation set of 100 sentences, similar to but distinct from the ones in the training sets. Following Google's recommendations, we configured the bots using Hybrid match mode, which is recommended for bots with few training examples for each intent. For the entities, we kept the default settings of allowing synonyms and automated expansion.

	Accuracy	
	Intent detection	**Slot filling**
A1	56%	31%
A2	84%	78%
Difference A1 - A2	28% (1.5x better)	47% (2.5x better)

We observed a significant improvement in accuracy for both intent detection and slot filling (intent + entities). A2 (the bot trained with the automatically extended training set) shows an improvement in accuracy of up to 1.5 times compared to that of A1 (the bot trained with the hand-tagged training set) for intent detection, and an even greater improvement of 2.5x for slot filling.

The second test (B) is very similar to the first one. The only difference is the number of sentences used in the training and evaluation sets. In this case, the first bot (B1) was trained with a hand-tagged training set of 50 sentences (10 per intent). Using those sentences as input, our Bitext Artificial Training Data service generated 798 sentences which, together with the 50 sentences from bot B1, were used to train the second bot B2 (with around 170 sentences per intent). We used the same 100 evaluation sentences from test A as the evaluation set.

	Accuracy	
	Intent detection	**Slot filling**
B1	64%	63%
B2	95%	90%
Difference B1 - B2	31% (1.5x better)	27% (1.4x better)

We also observed a significant improvement in test B, even more so than in test A, reaching at least 90% accuracy in both intent detection and slot filling in B2.

In summary, the Bitext Artificial Training Data service lets you create big training sets with no effort. If you only want to write one or two sentences per intent, our service is able to generate the rest of variants needed to go from really poor results to great accuracy. And even if you want to write tens of variants per intent, our service will also significantly increase the accuracy of your model, obtaining really good results. We have carried out these tests with Dialogflow, but our conclusions are relevant for ML-based bot platforms in general. We can conclude that our Artificial Training Data service is able to drastically improve the results of bot platforms that are highly dependent on training data.

Appendix

“Hybrid" match mode (default): accuracy increases when training datasets grow significantly, from 2-10 to 100+ sentences per intent with artificial data.

A1 (2-3 sentences per intent) vs A2 (around 80 sentences per intent):

- Intent detection: 56% vs 84% - increases
- Slot filling: 31% vs 78% - increases

B1 (10 sentences per intent) vs B2 (around 170 sentences per intent):

- Intent detection: 64% vs 95% - increases
- Slot filling: 63% vs 90% - increases