AlterNlu Documentation

Release v1.0.0-beta

Kontiki AI

May 06, 2019

Contents

A good-quality training dataset is the most critical aspect for building a robust and intelligent chatbot. Everything depends upon the efficiency of the training dataset for conversational agents to understand and respond, based on the user intent and context.

We understand that to revert a valid response to the user, each query has to first pass through a model. **And, to let our customers build, manage and analyse the training dataset, we developed** Alter NLU.

Alter NLU is an open source tool to train AI-based conversational agents, powered by deep learning. It is conceptualized by developers, for developers, enabling them to build high-quality dataset for chatbots of any domain.

Note: Alter NLU is programmed to maximize output with relatively less training data

This guide is segregated into 2 parts -

- Alter NLU Console
- Alter NLU Engine

This helps in developing a good-quality training dataset and building a NLU model for AI- based conversational agents.

CHAPTER 1

Understanding Alter NLU Console

The main focus of the Alter NLU Console is to get rid of the struggles involved in building a stable and good-quality training dataset. To achieve the same, it is segregated into 3 parts:

1.1 Understanding Alter NLU Console

1.1.1 Intents

Intents are what you expect users to say and what are their intentions. It depicts the gist of the expression that the user says or in simple terms what the user probably meant to say.

For Example :

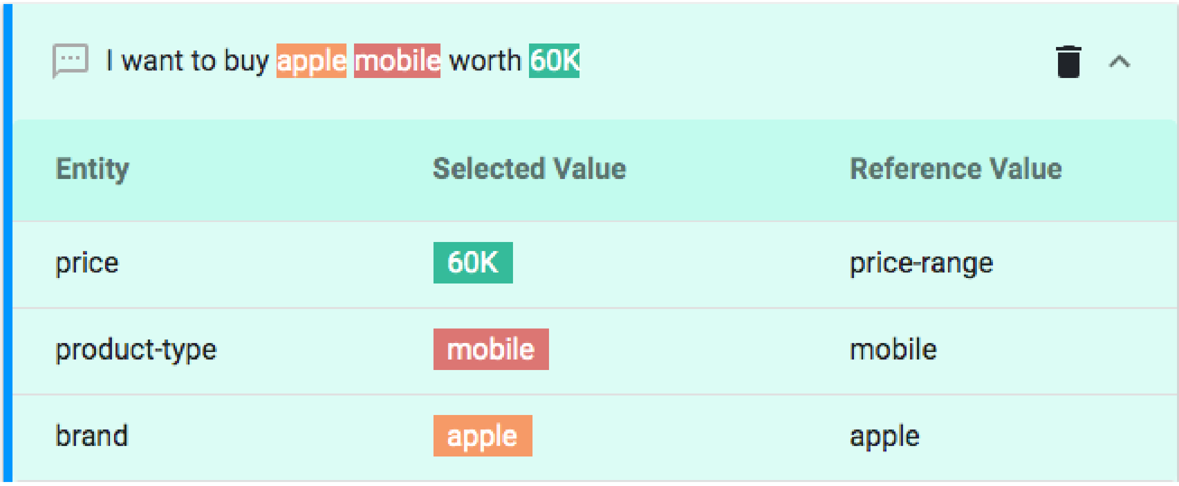
> If the user types : *I want to buy apple mobile worth 60k*
>
> Here the user intent to buy the specified product with the mentioned specifications. We can infer the intent of the user query as "search-product", which is a user-defined term.

There can be 2 varieties of intents:

- First, which do not contain any entities (explained below), but are simple and direct queries. For example "greet", "exit" and other light conversational intents

  ```
  "Hey there" or  "Hi".
  ```

- Second, which can contain multiple or single entities in the training query which we want to extract from the user query as demonstrated in the image below:



I want to buy apple mobile worth 60K

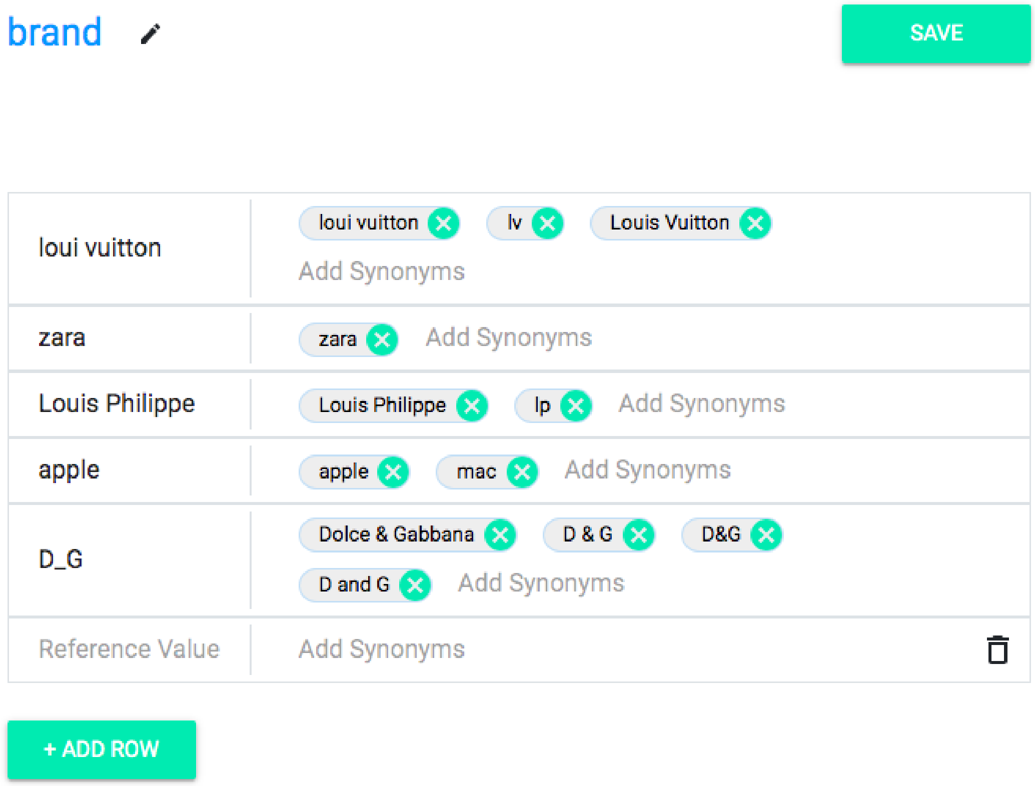
Entity	Selected Value	Reference Value
price	60K	price-range
product-type	mobile	mobile
brand	apple	apple

1.1.2 Entities

Entities describe the piece of information you would want to extract from the expressions/messages of the user.

Entities can consist of single or multiple words. For example "mobile" and "mobile cover" are 2 different entities.

Like in the image above, the "price", "product-type" and "brand" are the 3 entities in our Alter NLU console that will be tagged to their respective "Reference Value" and "Synonyms" values.



brand

loui vuitton	loui vuitton lv Louis Vuitton Add Synonyms
zara	zara Add Synonyms
Louis Philippe	Louis Philippe lp Add Synonyms
apple	apple mac Add Synonyms
D_G	Dolce & Gabbana D & G D&G D and G Add Synonyms
Reference Value	Add Synonyms

Reference Value

It is the convenient representation of the whole set of synonyms (explained below). It can be in the form of a unique_id, abbreviations, initials, shortforms, etc. according to the developers' convenience.

For example, in the last synonym set "Dolce & Gabbana" in the image above is conveniently referred as "D_G".

Synonyms

It is a set of words or phrases having similar meaning, mapped against each reference value.

Selected Value

It is the part of the sentence we highlight to tag against a particular entity. Each "Selected Value" in the Alter NLU console has a different color code based on the entity tagged.

1.1.3 Reports

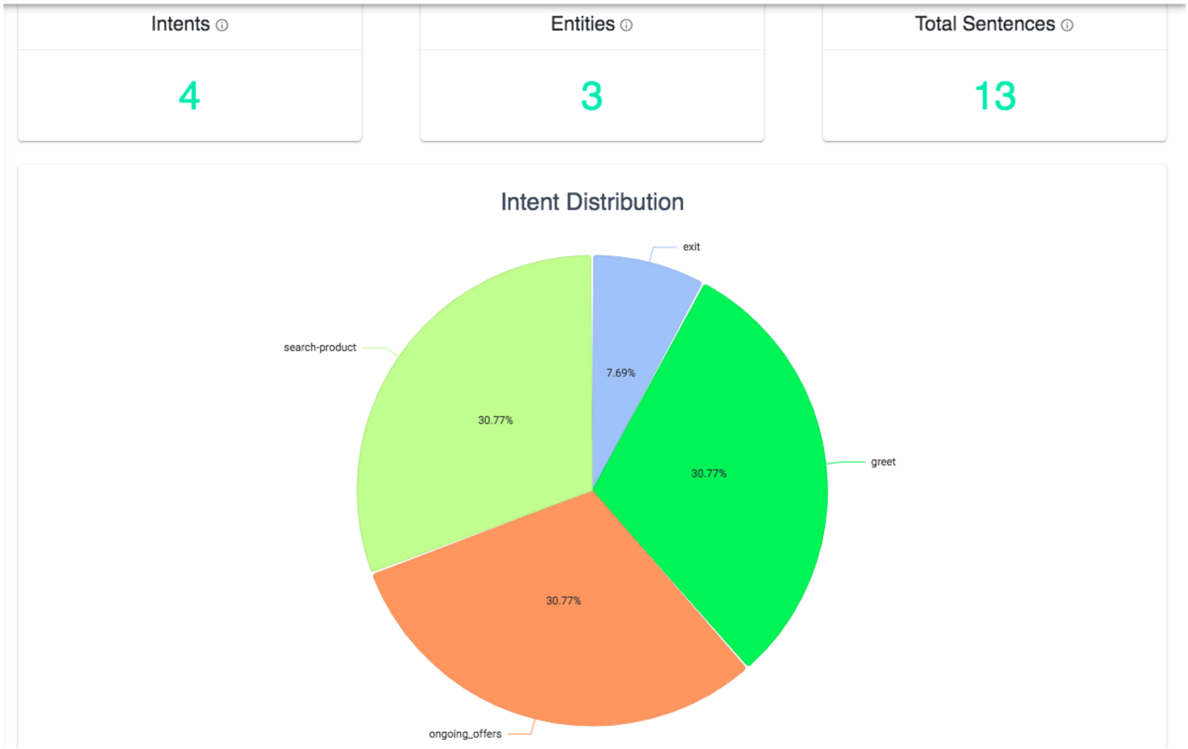
Chatbot training is an ongoing process that should get better at every successive stage. With each improvement, the trainer/developer should have a clearer understanding of the changes made.

> *The section provides NLP-based real-time 'Reports', where we list out the health of the training dataset created by the user, alert if there are issues and recommend ways to make it better.*

Here are the list of items that get covered under Reports:

- **Intent Distribution:** It represents that number of intents created and the numerical proportions for the number of relevant sentences present in each of the intents and their respective percentages.

  Note: The relevant sentences are those that contribute to building better a NLU model.

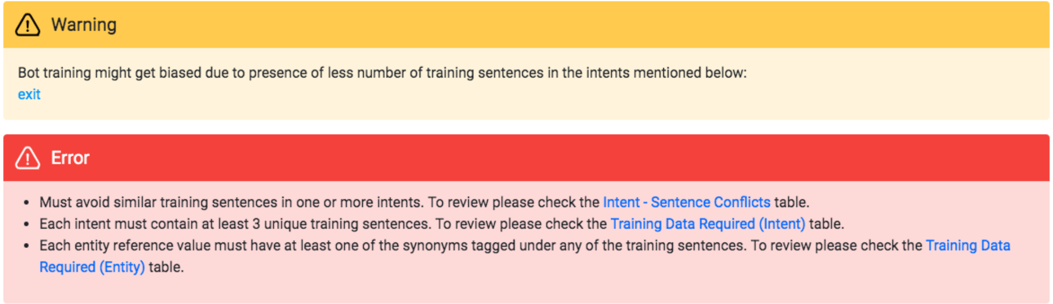
  

  Intents ⓘ 4

  Entities ⓘ 3

  Total Sentences ⓘ 13

- **Figuring out the intents that require more training sentences:** It reports the specific intents that have less number of training sentences than the threshold set i.e 3 relevant sentence per intent. Also, it notifies the user with the name of the intents lacking enough training queries in comparison to the other intents in the bot.



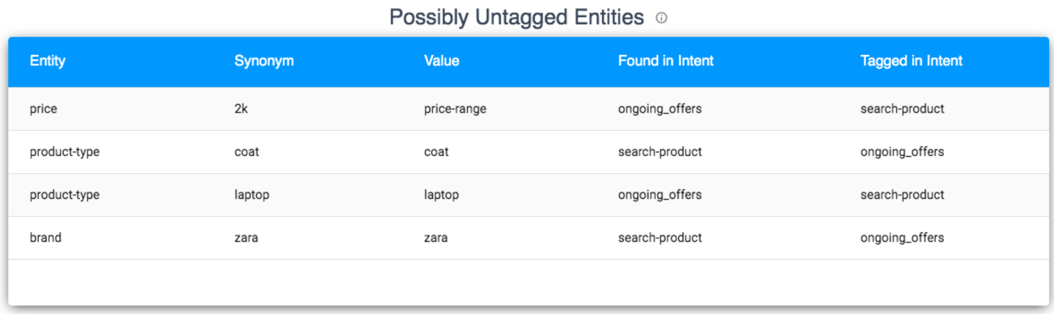
Warning

Bot training might get biased due to presence of less number of training sentences in the intents mentioned below:
exit

Error

- Must avoid similar training sentences in one or more intents. To review please check the Intent - Sentence Conflicts table.
- Each intent must contain at least 3 unique training sentences. To review please check the Training Data Required (Intent) table.
- Each entity reference value must have at least one of the synonyms tagged under any of the training sentences. To review please check the Training Data Required (Entity) table.

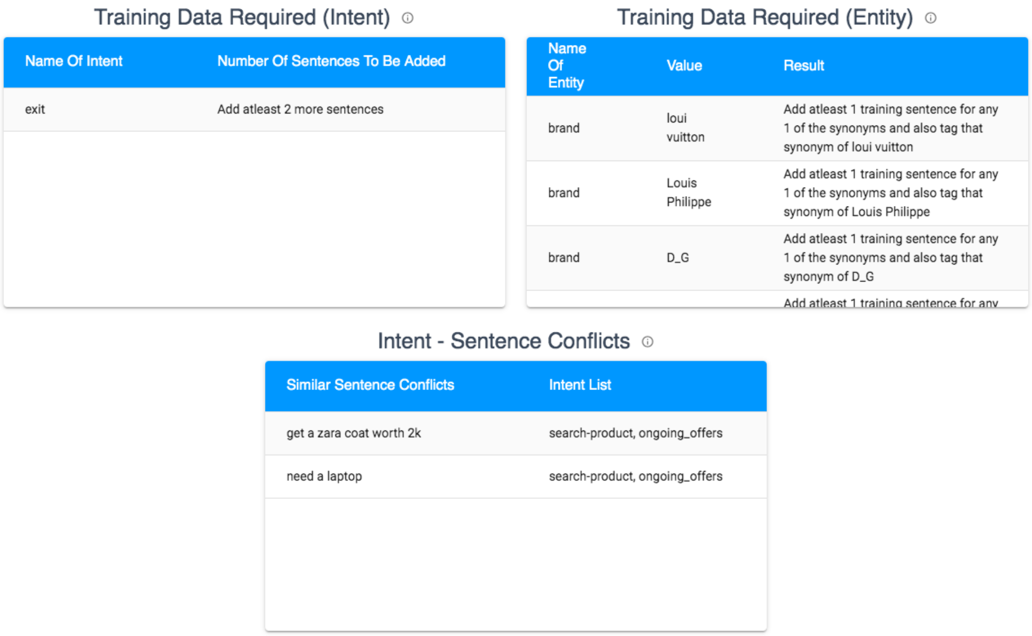
- **Examining the training dataset to extract the untagged entities:** Lists out keywords which have been tagged with an entity in intent but, the same keyword is untagged in the training sentence of another intent. It also notifies the user that they might have skipped tagging the keyword as an entity in the other intent mentioned.



Possibly Untagged Entities ⓘ

Entity	Synonym	Value	Found in Intent	Tagged in Intent
price	2k	price-range	ongoing_offers	search-product
product-type	coat	coat	search-product	ongoing_offers
product-type	laptop	laptop	ongoing_offers	search-product
brand	zara	zara	search-product	ongoing_offers

- **Listing out the limitations in the entity section:** It reports about the name of the entities that have been defined but have not been used by the user to form training queries in the intent section. It also reports when the user might have mistakenly deleted the entity from the intent section but forgotten to delete the same from the entity section
- **Capturing repetition of training sentence:** It informs about the training sentence(s) that the user might have added in multiple intents by mistake. The console alerts this to the user with an error message at the top of the reports section.



Training Data Required (Intent) ⓘ

Name Of Intent	Number Of Sentences To Be Added
exit	Add atleast 2 more sentences

Training Data Required (Entity) ⓘ

Name Of Entity	Value	Result
brand	loui vuitton	Add atleast 1 training sentence for any 1 of the synonyms and also tag that synonym of loui vuitton
brand	Louis Philippe	Add atleast 1 training sentence for any 1 of the synonyms and also tag that synonym of Louis Philippe
brand	D_G	Add atleast 1 training sentence for any 1 of the synonyms and also tag that synonym of D_G
		Add atleast 1 training sentence for any

Intent - Sentence Conflicts ⓘ

Similar Sentence Conflicts	Intent List
get a zara coat worth 2k	search-product, ongoing_offers
need a laptop	search-product, ongoing_offers

Note: For good results, the console requires a minimum of 3 relevant training queries per intent. Also, Alter NLU console needs only a single synonym tagged for each entity reference value in training queries of the intent section. So, all you need to do is add 1 training query in the intent section containing any one of the synonyms per reference value and your bot is good to go.

For instance, we need to train for 1 user query containing synonym for “D_G” reference value, like - *I want a Dolce & Gabbana bag* and rest will be handled automatically.

CHAPTER 2

Alter NLU Console Features and Data Manipulation

From interactive user-interface to getting real-time report of training dataset. Alter NLU provides all the necessary features that are required to create a robust dataset. Click on the link below to have a brief overview of what all we include in this open source tool to train AI based conversational agents.

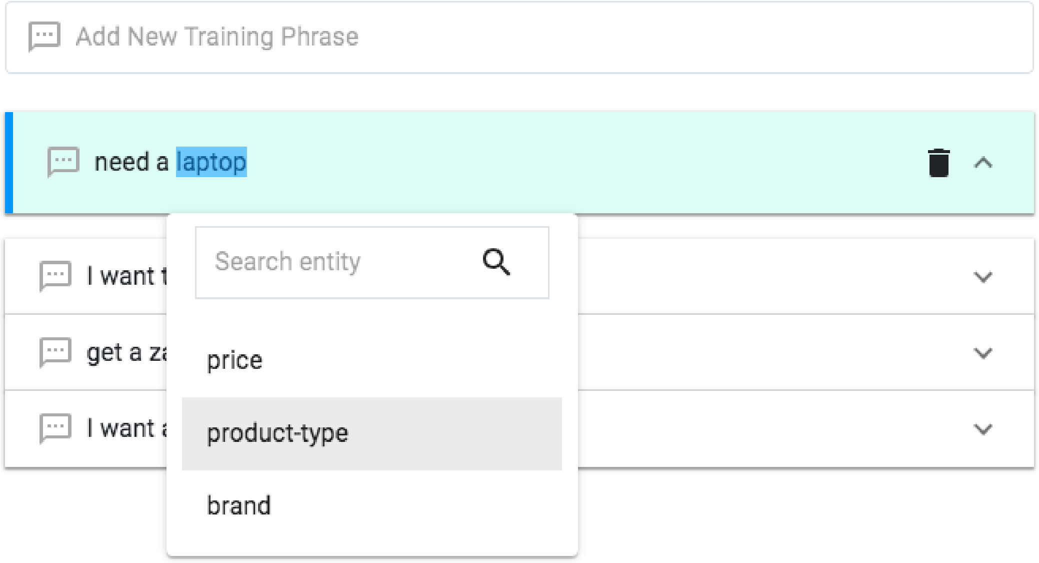
2.1 Alter NLU Console Features and Data Manipulation

2.1.1 Interactive UI to Build and Manage Training Data

Intent : Create, Modify, Delete intent and intent queries.

We give filter functionalities for intent names and intent-specific training sentences to help the user modify the data quickly. We also have a search functionality that enables users to filter out the list.

For user convenience, we allow drop down search for both "Selected Value" and "Reference Value".



Entity : Create, Modify, Delete entity and entity-specific reference value and synonyms.

We give filter functionalities for entity name, reference value and entity-specific synonyms to help the user modify the data quickly.

> **Note:** Whenever an entity is tagged in the sentence, it is automatically mapped in the synonyms of the corresponding reference value.
>
> Also, if you change the 'Reference Value' in the entity section the same will be reflected dynamically in the intent section, and vice versa.

2.1.2 Get Real-Time Report of Training Data if There's Any Issue

A dedicated reports page which gives you the insights of all the warnings and errors that might affect the training of the bot or make it perform less efficiently.

- Presenting Intent Distribution in the form of a pie chart
- Figuring out the intents that require more training sentences
- Listing out the limitations in the entity section
- Examining the training dataset to extract untagged entities
- Capturing repetition of training sentence

2.1.3 Download the Training Dataset in 2 Formats - Alter NLU & RASA NLU

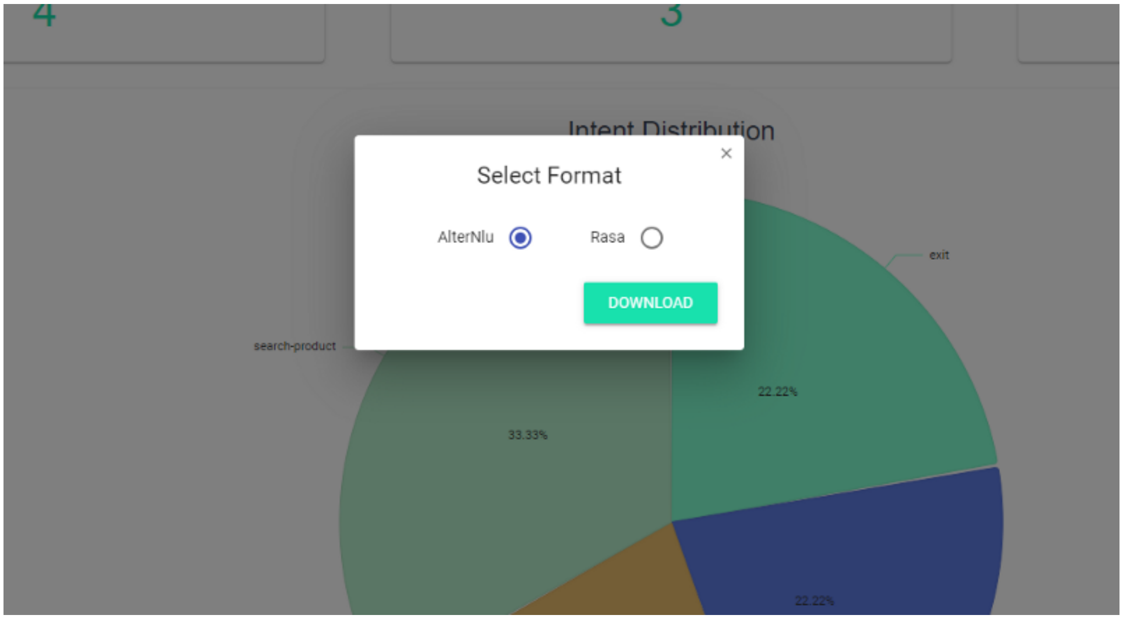
The download button will get activated only when user meets the below threshold requirements -

1. Each intent must contain at least 3 relevant training sentences.
2. Each entity reference value must have at least one of the synonyms tagged in any one of the intents' training sentences.

3. There should never be multiple intents containing same training sentence(s).
4. The number of total intents must be at least 2 to allow downloading of training data.

Once you have rectified all the errors, you will be able to download the dataset JSON in both—the Alter NLU and the RASA format.

> **Note:** If you are using RASA NLU, you can quickly create the dataset using Alter NLU Console and Download it in RASA NLU format. We have updated our console for hassle free data creation which is less prone to mistakes.


Select Format

AlterNlu ● Rasa ○

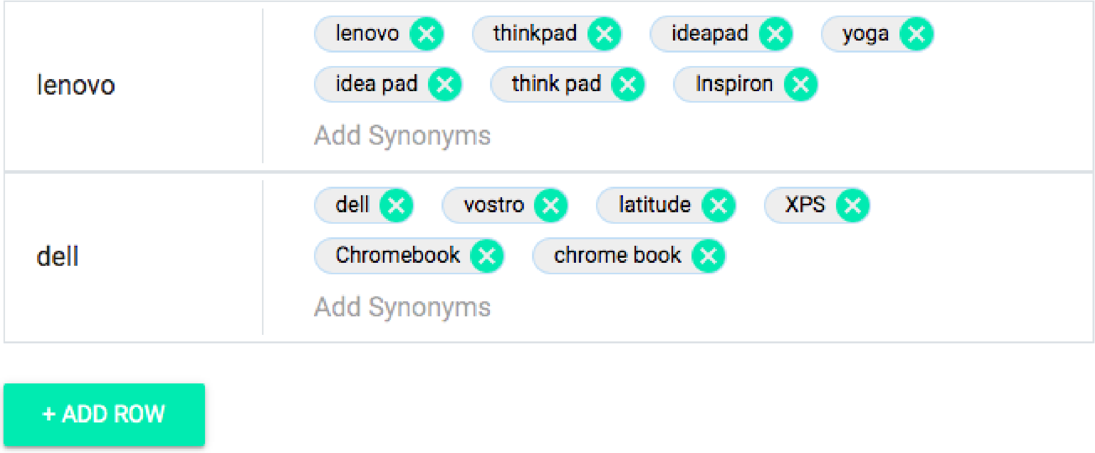
DOWNLOAD

2.1.4 Data Manipulation

To maintain the dataset standards, we apply dynamic algorithms to perform data manipulation efficiently. Below is an example that illustrates how we are manipulating your training data for better accuracy.

> **Note:** Any modification made in the entity section is altered dynamically in the sentences of the intent section and vice versa.

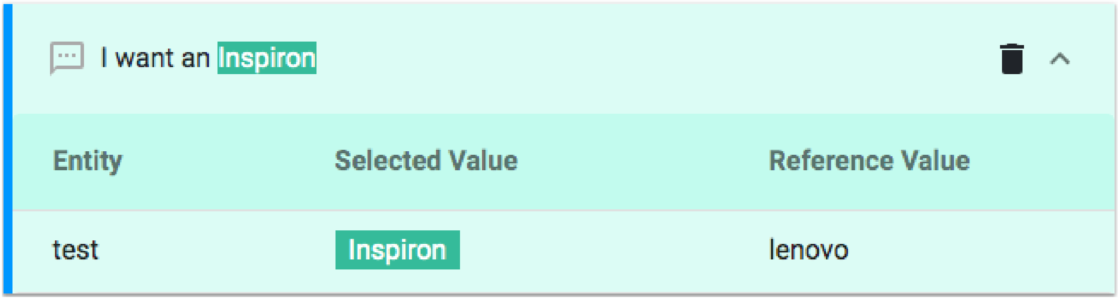
Let us suppose - an entity “brand” which has the below data:



lenovo	lenovo thinkpad ideapad yoga idea pad think pad Inspiron Add Synonyms
dell	dell vostro latitude XPS Chromebook chrome book Add Synonyms

From the image above we can make out that Reference Value i.e "lenovo" has synonyms - `["inspiron", "thinkpad"]` etc, while the other entry is "dell" which holds `["vostro", "chromebook"]` etc as synonyms.

Now, in the intent section, I train for the phrase - "I want an Inspiron". And for other similar phrases, I tag the word "Inspiron" with "lenovo" reference value.



I want an Inspiron

Entity	Selected Value	Reference Value
test	Inspiron	lenovo

Later, while examining my created entities, I realize that I have added "Inspiron", which is a variant of "dell" to "lenovo". Therefore, from the entity section I delete the synonym "Inspiron" from "lenovo" and add it to "dell" reference value. Now, our code dynamically judges the modification made and update the "Reference Value" to "dell" in all the sentences present in the intent section.

CHAPTER 3

Alter NLU Engine

Here, we will be covering the Alter NLU GitHub Repo and how to create a NLU model for your chatbot/AI Assistant. The guide provides a detailed explanation about the benefits of using the Alter NLU Engine with a query example.

3.1 Alter NLU Engine

3.1.1 Installation

Go to this GitHub Repo and follow the steps in the “Setting Up” section of the README.MD file.

3.1.2 Rest API

Rest API supported for both the below training and parsing queries.

- REST API training

```
URL     : http://<ip_address>:5001/train
Method  : POST
Headers : {
    Accept : application/json,
    Content-Type : application/json
}
Body    : Content of Training Data JSON file downloaded from Alter NLU Console
```

  or

```
curl -H "Content-Type: application/json" --data @<training_data_json_file_path>␣
↪http://<ip_address>:5001/train
```

- Rest API parse query

```
URL    : http://<ip_address>:5001/parse
Method : POST
Headers: {
    Accept : application/json,
    Content-Type : application/json
}
Body   : {"text": "<user_query>"}
```

3.1.3 v1.0.0-beta: The Engineering Involved

Intent Model

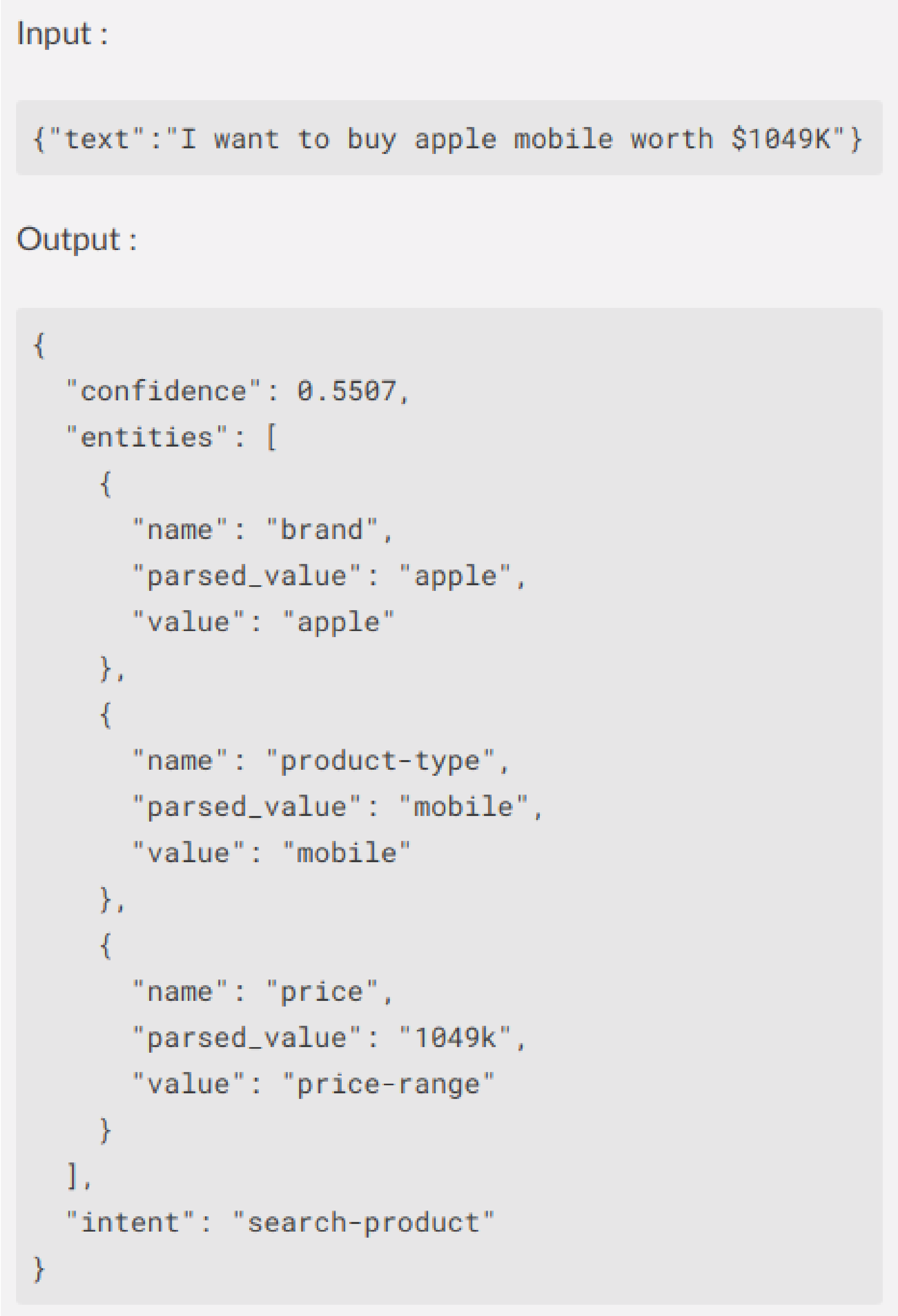
We have used **Convolutional Neural Networks (CNN)** based model to capture the intent. Further, the use of **custom validation algorithm** and **matthews correlation coefficient** as accuracy metrics makes the intent model robust.

Entity Model

In this version, we have replaced the previous Flashtext and FuzzyWuzzy based entity extraction method with a **CRF based Entity Recognition model**.

3.1.4 Alter NLU Engine Response Format

A detailed explanation of the response and an example is given below:



Input :

```
{"text":"I want to buy apple mobile worth $1049K"}
```

Output :

```
{
  "confidence": 0.5507,
  "entities": [
    {
      "name": "brand",
      "parsed_value": "apple",
      "value": "apple"
    },
    {
      "name": "product-type",
      "parsed_value": "mobile",
      "value": "mobile"
    },
    {
      "name": "price",
      "parsed_value": "1049k",
      "value": "price-range"
    }
  ],
  "intent": "search-product"
}
```

- According to the context of the user query, the model successfully recognises the search product intent along with the confidence score.

- This model handles *out-of-vocabulary* words to some extent. The term *'out-of-vocabulary words'* refers to words that are not present in the training data of the chatbot. For instance, you trained in ALter NLU for the sentence:

  I want to purchase apple mobile worth 60k

  Now, take a look at the input JSON, formed based on the user query in the image above. Even though the **parsed_value**, "1049k" may not be present in your training dataset, the output recognises the entity accurately as "price".
- The **CRF model** helps is recognising the entity accurately, because it considers the sentence structure of the user query.
- The main goal of the **"parsed_value"** key in the response is to assist developers to directly use the key where needed. In the example above, the developer might need the exact value of entities such as "price", which is in the user query for further usage. In this case it is "1049k".
- Also, if you are an existing user of Alter NLU, it needs to be pointed out that the "category" key in the response has been renamed to "name".

CHAPTER 4

Building an E-commerce Chatbot

We have made the Alter NLU console easy for you to get started and build your own chatbot training dataset.

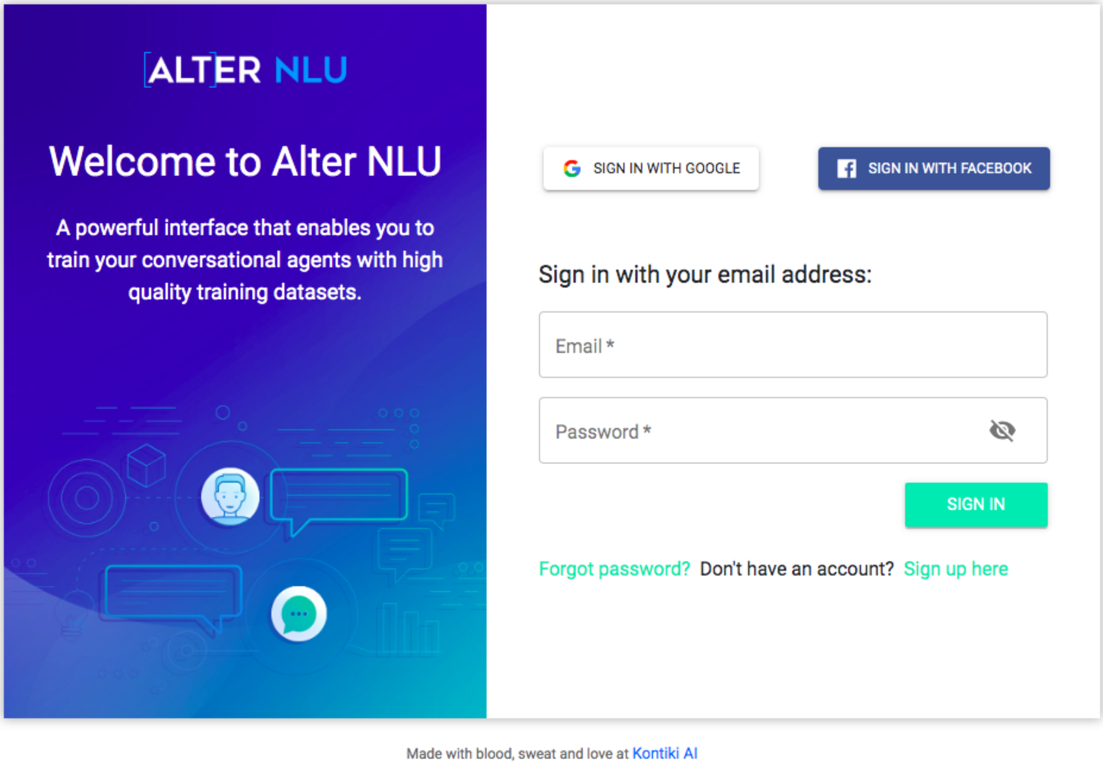
This step-by-step guide begins with a tutorial on creating your account and adding your first chatbot dataset. We also demonstrate how to manage and analyze customized training dataset in real-time.

In this guide, for the ease of reference and explanation, we are going to use e-commerce dataset as an example to show how a chatbot can be trained to purchase items.

4.1 Building an E-commerce Chatbot

The console is developed to handle multiple chatbot datasets within a single user login i.e you can add training data for any number of chatbots.

Start by creating your account from our Alter NLU console



Welcome to Alter NLU

A powerful interface that enables you to train your conversational agents with high quality training datasets.

Sign in with your email address:

Don't have an account?

Made with blood, sweat and love at Kontiki AI

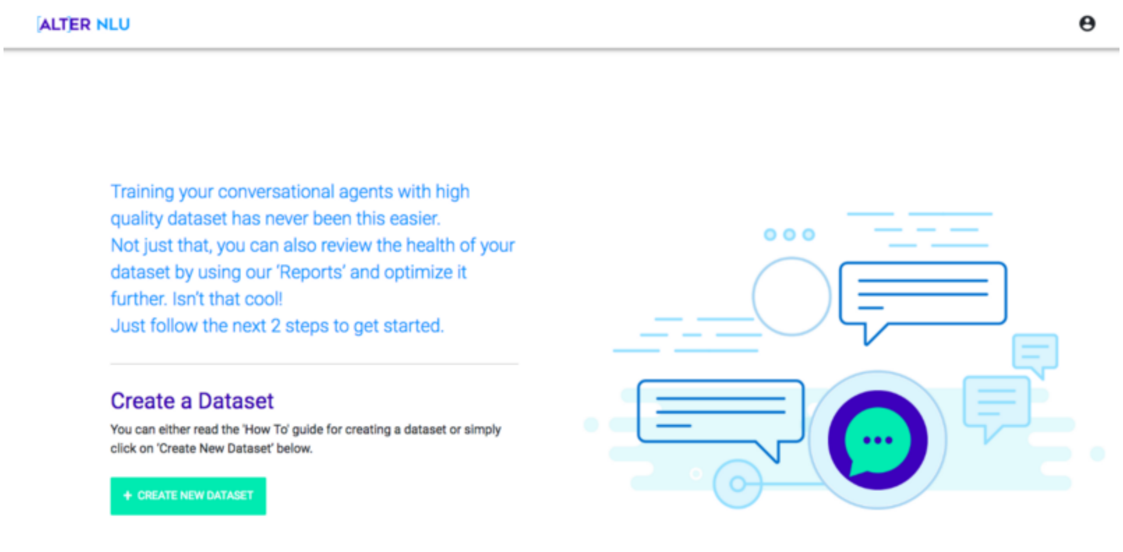
Upon clicking the 'Sign Up' button, an e-mail is sent to the ID you provide above. You can log in on the console once you verify your identity via the e-mail.

You can also register and login using social sites credentials of your Facebook or Google account.

Below in the documnet are the steps that will guide you in building your training dataset.

4.1.1 Creating Dataset

Upon successful login, you will be redirected to the "Create Dataset" page.



ALTER NLU

Training your conversational agents with high quality dataset has never been this easier.
Not just that, you can also review the health of your dataset by using our 'Reports' and optimize it further. Isn't that cool!
Just follow the next 2 steps to get started.

Create a Dataset

You can either read the 'How To' guide for creating a dataset or simply click on 'Create New Dataset' below.

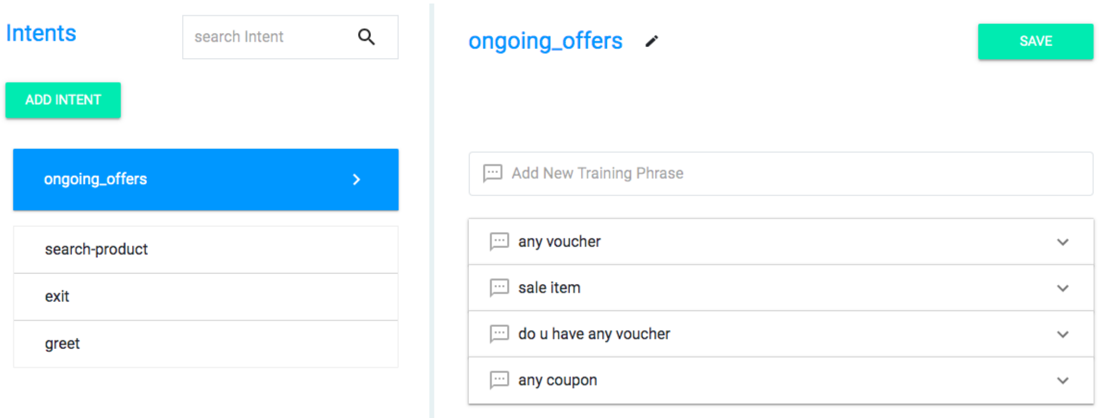
Get started by creating a new dataset, which requires a `bot name` and the `industry/vertical` that your bot belongs to. Here, we are going to name our bot as - "ecomm-bot" and the domain will be "E-commerce". Once you click on the "Add" button, the dataset gets created and you will be redirected to "Intent Page".

4.1.2 Building no Keyword Intents (Intents with no Entities):

The intents like "greet" and "exit" are the generic intents every bot should handle. Besides these intents, build other intents you plan to include in your chatbot. Like, I have included "ongoing_offers". Follow the steps below to create the "ongoing_offers" intent :

- Create an intent using the "ADD INTENT" option.
- Next, in "Add New Training Phrase" text area write simple user queries asking about the current sales, vouchers in our e-commerce chatbot. Example - 'any offers?', 'do you have any voucher', etc.
- Save the "ongoing_offers" intent using the "SAVE" button on the top right corner.

This intent will hold all the user queries asking about the current sales, vouchers in our e-commerce chatbot.



Intents: ongoing_offers, search-product, exit, greet

ongoing_offers

- any voucher
- sale item
- do u have any voucher
- any coupon

Remember to train the dataset with expressions that contain words like sales, vouchers, etc. This is because words will keep the "ongoing_offers" intent unique from other non-keyword intents.

4.1.3 Building Entities

The e-commerce chatbot should be trained to handle queries like:

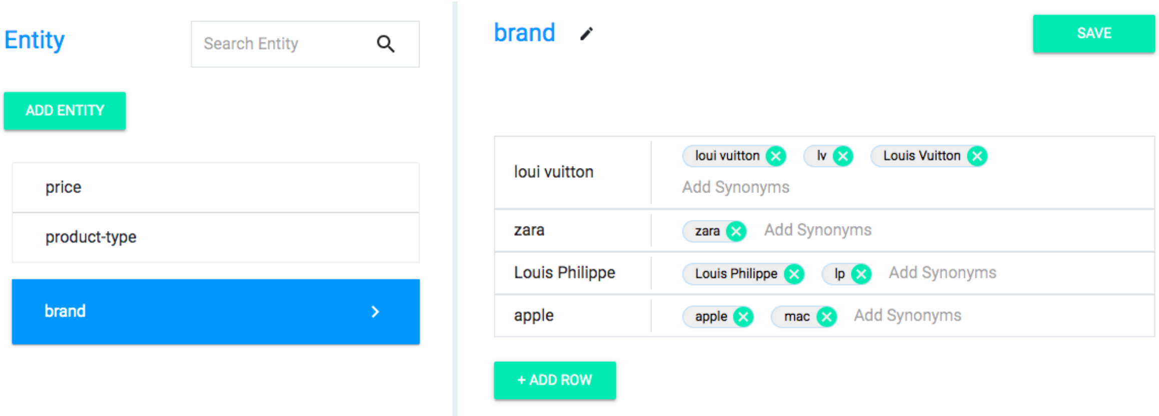
User says : I want to buy **apple mobile** worth **60K**.

So, I plan to create 3 entities that will extract the:

- brand
- product-type and
- price

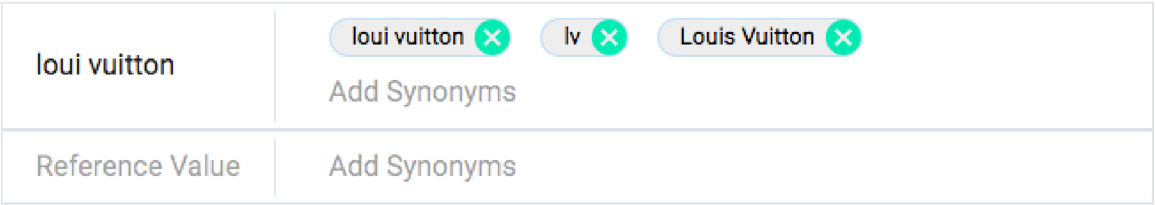
from the user query.

- Create an entity "brand" by using the "ADD ENTITY" option.
- The "brand" entity will contain "Reference Value" as the main brand name (like apple) as well as synonyms that the user may refer to a particular brand that your chatbot endorses.



loui vuitton	loui vuitton ⊗ lv ⊗ Louis Vuitton ⊗ Add Synonyms
zara	zara ⊗ Add Synonyms
Louis Philippe	Louis Philippe ⊗ lp ⊗ Add Synonyms
apple	apple ⊗ mac ⊗ Add Synonyms

- Similarly create other entities, add the 'Reference Value' and its synonyms. Like a user can write 'loui vuitton' (our Reference Value) as 'lv' or 'Louis Vuitton' (the synonymns for 'loui vuitton').



loui vuitton	loui vuitton ⊗ lv ⊗ Louis Vuitton ⊗ Add Synonyms
Reference Value	Add Synonyms

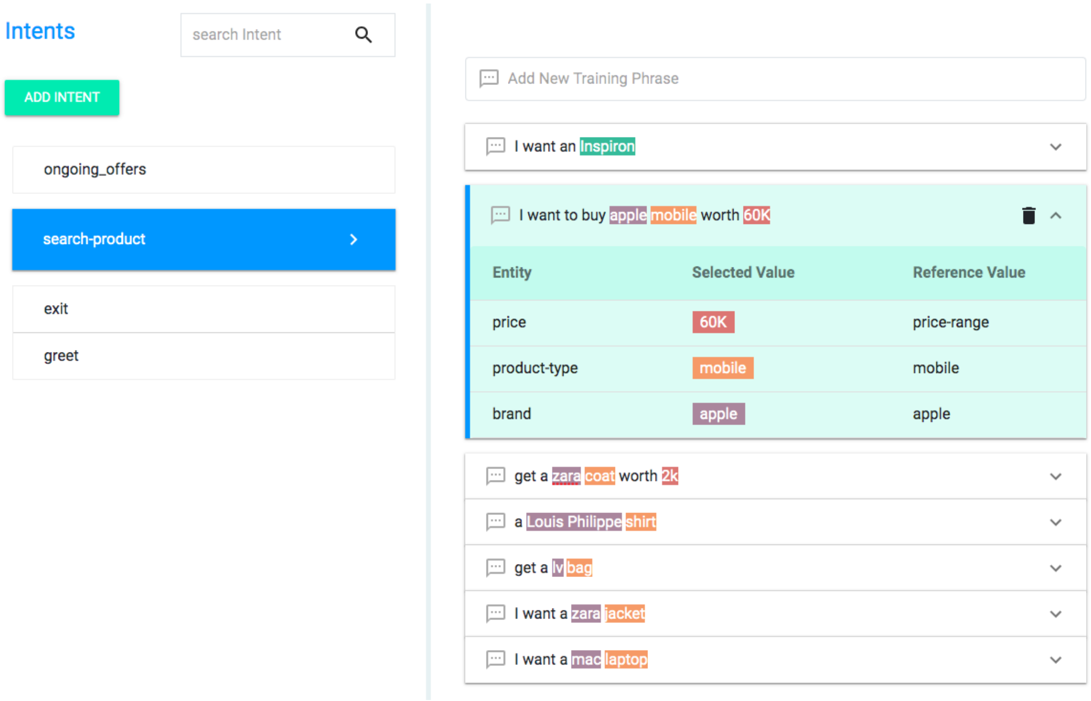
- Similarly, create values and synonyms for the other fields and finally, save your dataset.

4.1.4 Building Intents with Entities

For queries as stated in the above section, dataset should have an intent that stores all possible user queries from which the bot should be extracting the entities.

Create an intent with the name "search-product" and go to the training phrase section of the intent and start writing the expected user queries.

For instance, "I want to buy **apple mobile** worth **60K**". From this text, tag the information you want to extract and work upon.



I want an Inspiron

I want to buy apple mobile worth 60K

Entity	Selected Value	Reference Value
price	60K	price-range
product-type	mobile	mobile
brand	apple	apple

get a zara coat worth 2k

a Louis Philippe shirt

get a lv bag

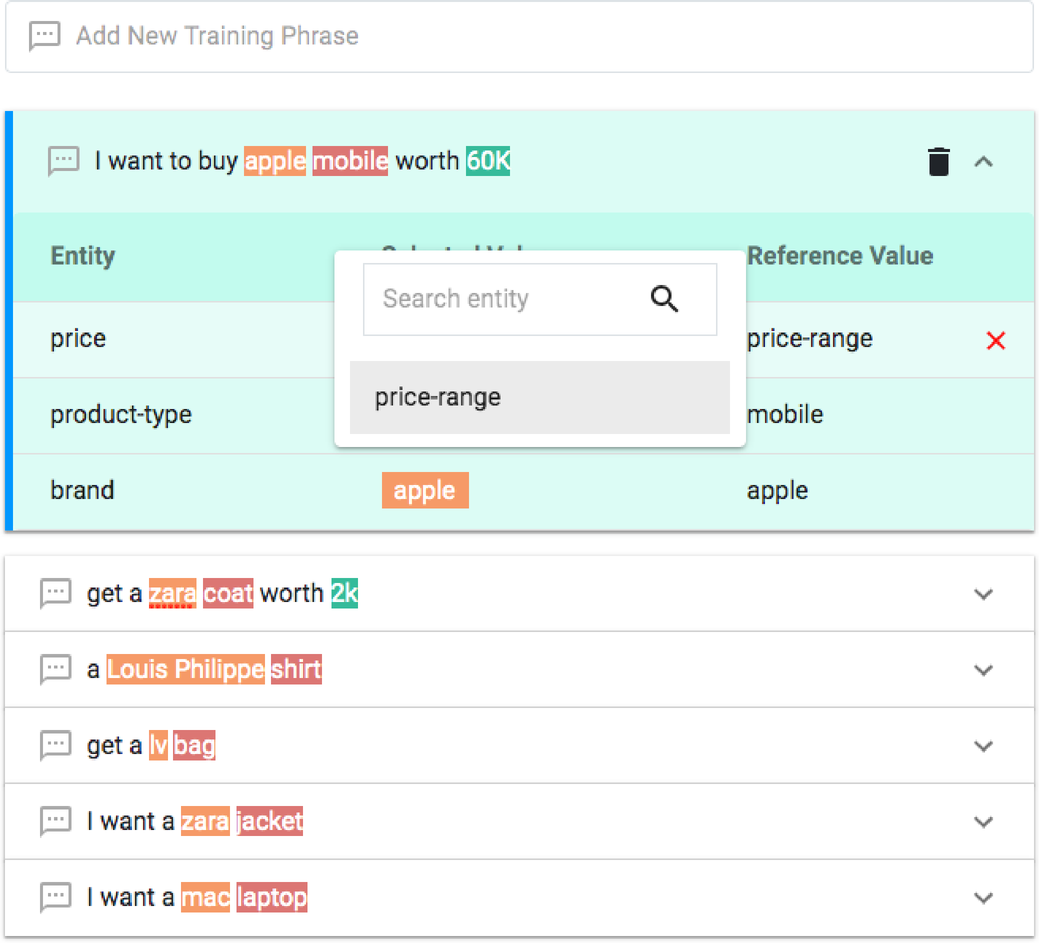
I want a zara jacket

I want a mac laptop

For the ease of developers, we have built the console in a manner that each ‘Selected Value’ in the intent section can be linked to a ‘Reference Value’ of your choice.

Like in the images below, you can see in the intent section:

Selected Value : 60k, 2k both have same ‘Reference Value’ i.e price-range



Add New Training Phrase

I want to buy apple mobile worth 60K

Entity		Reference Value
price		price-range
product-type		mobile
brand	apple	apple

Search entity

price-range

get a zara coat worth 2k

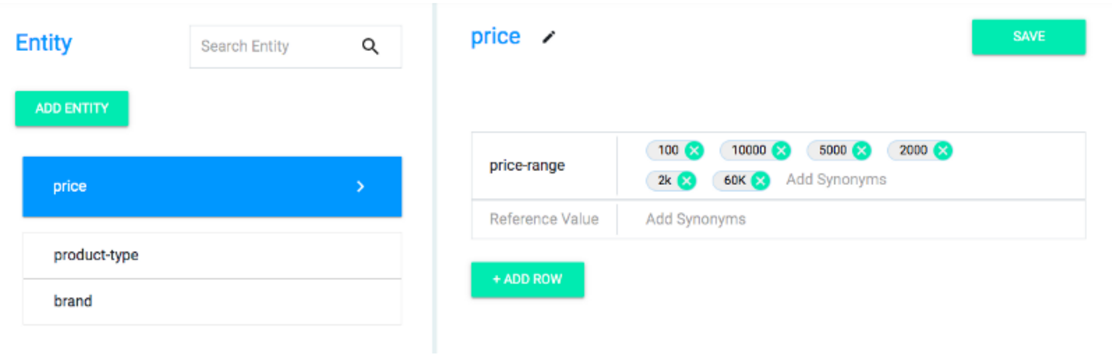
a Louis Philippe shirt

get a lv bag

I want a zara jacket

I want a mac laptop

In the entity section, every price-range example is defined in the same "Reference Value"



price-range	100 10000 5000 2000 2k 60K Add Synonyms
Reference Value	Add Synonyms

Note: If you change the 'Reference Value' in the entity section, the same will be reflected dynamically in the intent section and vice versa.

4.1.5 Analysing Loopholes in the dataset: The Report Section

Once you are done building the dataset, move to the Report Section which will analyse your dataset for all intents and entities in real-time and notify the errors and warnings that need to be addressed for the accuracy of the chatbot's response.

Click on the tool tip icon next to each of the section title for recommendations. Below listed are the key functions that form the Report:

- Illustrating Intent Distribution in the form of a pie chart.
- Analysing the Training Data Required by intents and entities to achieve accuracy.
- Pointing out Possibly Untagged Entities to inform about keywords that have been tagged as an entity in the intent, but the same keyword occurs untagged in the training sentence of another intent.
- Intent - Sentence conflicts table alerts about the training sentence(s) you may have added in multiple intents by mistake.
- Handling training bias by highlighting the name of intents lacking enough training expressions when compared with other intents.

Once you have rectified all the errors, you will be able to download the dataset JSON in both—the Alter NLU or the RASA format.

Note: If you are using RASA NLU, you can quickly create the dataset using Alter NLU Console and Download it in RASA NLU format. We have updated our console for hassle-free data creation that is less prone to mistakes.

4.1.6 Build Your Bot

Go to Git Repository from the link below:

https://github.com/Kontikilabs/alter-nlu/tree/v1.0.0-beta

Next, go through the README.MD file and start executing the steps as mentioned.