bitext we help AI understand humans

Bitext Pre-Built Training Data for Intent Detection

Introduction

Building effective conversational agents requires large amounts of training data. Producing this data manually is an expensive, time-consuming and error-prone process which does not scale. Platform providers usually do not have the infrastructure required to tackle the wide range of verticals, languages and locales that their large clients need to handle, while clients rarely have the expertise necessary to collect and annotate their data, and outsourcing the task is complicated by the fact that the data often contains sensitive data that cannot be exposed to third parties.

Bitext Pre-Built Training Data for Intent Detection accelerates the deployment of new IVR, case routing and chatbot models by bootstrapping the training process with pre-packaged synthetic data for a wide range of verticals, so customers can have a multilingual system up and running in a few hours.

Our ready-to-use synthetic training datasets cover the most common intents for each vertical and include a large number of example utterances for each intent, with optional entity/slot annotations for each utterance. The datasets offer users a way to quickly and easily deploy conversational agents on any platform, without having to understand the technical details of how to train optimal AI models. By using synthetic data, our templates offer scalability, modularity, customizability, consistency and high coverage.

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

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

Methodology

We employ a scalable and data-driven linguist-in-the-loop methodology. We begin by collecting large volumes of text from domain-specific public data sources such as FAQs, knowledge bases and technical documentation.

We then apply our Deep Parsing technology to automatically extract the most frequent actions and objects that appear in those texts. This results in a knowledge graph that captures the semantic structure of the vertical, which is then curated by computational linguists to identify synonyms and to ensure consistency and completeness. Actions are grouped into categories and intents, and the intent structure is then validated against FAQs and with domain experts.

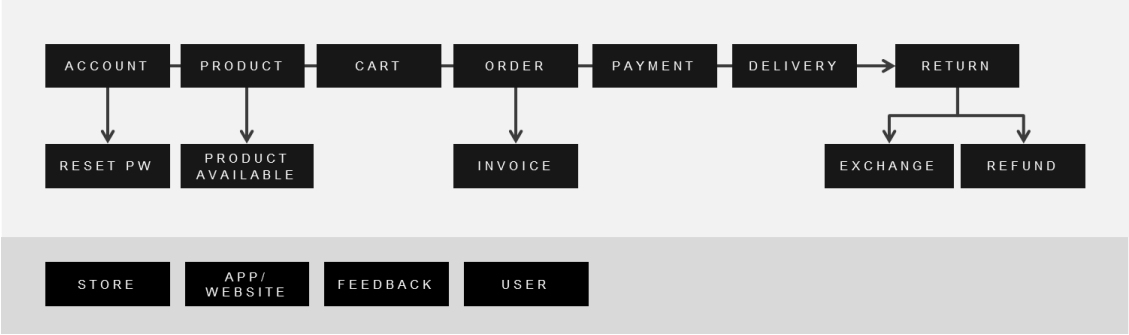

Figure 1. Retail Vertical Map

Finally, the linguistic structure of each intent is defined, together with the applicable frame types which allow our Natural Language Generation (NLG) technology to generate utterances which are predictable and consistent semantic variations of each intent request. This approach provides a measurable improvement to NLU performance: benchmarks comparing a manual baseline with our synthetic data show a >30% increase in intent detection and slot filling accuracy across multiple platforms.

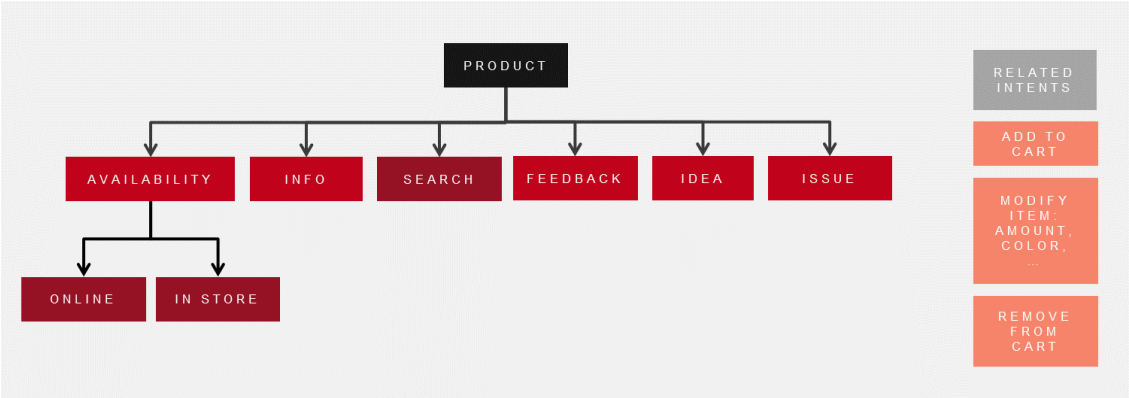

Figure 2. Product Category Map

Our methodology and tools allow us to easily customize and adapt the datasets to changing needs, including new intents, corporate terminology, language registers, new regions, markets and languages. With each change, the data is automatically regenerated, allowing for continuous improvement in a scalable fashion.

Contents

Each vertical template contains 20 to 100 common intents, 500 to 5,000 utterances per intent and entity/slot annotations for each utterance. The data is organized into one core dataset and several optional advance module datasets.

The core dataset contains:

- Initial definition based on common linguistic phenomena, including:
  - lexical: *item*, *product*; *modify*, *change*…
  - syntactic: *I want to*…, / *how can I*…
  - plus morphology, coordination…
- Structured according to specific language use:
  - lexical: vertical terminology
  - syntactic: vertical audience language

The advanced module datasets cover the following additional linguistic parameters:

- Politeness: *could you…, please?*
- Expanded abbreviations: *I'd like → I would like…*
- Colloquial: *i wanna…*
- Indirect: *ask an agent to…*
- Offensive: *exchange a f***ing product*
- Regional: *cancelling/canceling, basket/cart*

Verticals

The following verticals are currently available:

- Automotive
- Retail Banking
- Education
- Events & Ticketing
- Field Service
- Healthcare
- Hospitality
- Insurance
- Legal Services
- Manufacturing
- Media Streaming
- Mortgages & Loans
- Moving & Storage
- Real Estate / Construction
- Restaurant & Bar Chains
- Retail / E-commerce
- Telecommunications
- Travel
- Utilities
- Wealth Management

Languages

Pre-Built Training Data for Intent Detection is currently available in English and Spanish.

Our NLG technology is already available for German, French, Italian, Danish, Portuguese, Dutch and Swedish. Templates in these languages will be available in Q1 of 2020. Support for Turkish, Polish, Chinese, Japanese and Korean will be available in Q3 of 2020.

Contact Us

North America
Daniel Benito
dbenito@bitext.com

EMEA
David Fernández
david.rubi@bitext.com

Annex

Data Available Now (Q1 2020)

Industry	Language	Intents	Core Utterances	Total Utterances
Automotive	English	52	35233	335987
Retail Banking	English	26	24925	156963
Education	English	37	39560	407425
Events & Ticketing	English	25	59237	535403
Field Service	English	27	28335	267405
Healthcare	English	40	32968	320220
Hospitality	English	24	59122	585196
Insurance	English	38	46635	444162
Legal Services	English	29	23881	240504
Manufacturing	English	34	30055	317610
Media Streaming	English	24	37717	420051
Mortgages & Loans	English	39	43868	388751
Moving & Storage	English	29	26114	255661
Real Estate / Construction	English	28	23757	238659
Restaurant & Bar Chains	English	30	27489	261823
Retail / E-commerce	English	34	66772	656979
Telecommunications	English	26	22278	190274
Travel	English	33	43452	447606
Utilities	English	21	27482	271672
Wealth Management	English	24	20211	97516
Total	English		**70,6301**	**6,781,308**

Industry	Language	Intents	Core Utterances	Total Utterances
Automotive	Spanish	43	67943	1150901
Retail Banking	Spanish	24	25662	374238
Education	Spanish	39	45084	723125
Events & Ticketing	Spanish	25	235439	3627043
Field Service	Spanish	27	63881	1087319
Healthcare	Spanish	38	73942	1203086
Hospitality	Spanish	25	166580	3019366
Insurance	Spanish	36	73783	1404373
Legal Services	Spanish	27	31732	512516
Manufacturing	Spanish	34	23910	456760
Media Streaming	Spanish	23	42781	671009
Mortgages & Loans	Spanish	39	120227	2188130
Moving & Storage	Spanish	29	27975	437805
Real Estate / Construction	Spanish	29	48487	728168
Restaurant & Bar Chains	Spanish	30	51744	921056
Retail / E-commerce	Spanish	30	63563	940029
Telecommunications	Spanish	26	161847	2406239
Travel	Spanish	31	66136	1066449
Utilities	Spanish	21	26681	414800
Wealth Management	Spanish	21	54974	843743
Total	Spanish		**1,472,371**	**24,176,155**

Note: the *Core Utterances* figures represent the number of utterances in each industry's core dataset (covering lexical, syntactic, morphological and coordination phenomena), while the *Total Utterances* figures represent the total number of expanded variants in the advanced module datasets (which also cover politeness, colloquial, offensive, regional…).

The Synthetic Training Data currently available is designed primarily for Customer Support Automation. The intents provided cover the most common types of customer support interactions, and the number of intents varies based on each industry's specific needs.