Bitext Natural Language Generation (NLG) to help chatbot training

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 (for example, “close the kitchen door”) you need to provide it with a large number of sentences that express that intent (such as “could you close the door of the kitchen” or “do you mind closing the kitchen door, please?”).

Until now, these large training corpora had to be generated manually, with one or more people writing many different sentences for each intent, vertical and language that needed to be recognized in your chatbot. This is a time-consuming task rather than a creative one, and it makes successful bot development very costly.

At Bitext, we solve this common problem with our proprietary Artificial Training Data technology (also called Natural Language Generation or Variant Generation), 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.

NLG procedure

The process of creating an NLG corpus usually comprises three steps: initial seed sentences, NLU (Natural Language Understanding) and NLG (the generation itself).

1. Initial seed sentences

The owner of the bot defines a few sentences (5-10 at most) for each of its intents. For example, for a “price request” intent, the sentences may be “what is the price of X?” and “How much does X cost?”. These sentences reflect the bot owner’s experience of their customers’ behavior, thus adding a perspective that can not be replaced by any machine. This is a very small effort compared with the usual effort required to train a chatbot (100-200 example sentences per intent).

2. NLU process

Bitext analyzes each one of the seed sentences with their NLU service, which identifies the intent and the main parts (or “slots”) of the sentence. For example, in the sentence “do you mind closing the kitchen door?” it identifies:

- “intent”: “close”
- “object”: “door”
- “place”: “kitchen”

It then passes these values on to the NLG process, to start generating all variants for that structure.

3. NLG process

The NLG process receives the information output by the previous NLU analysis and generates a number of sentences which the same meaning. These sentences vary in different aspects:

- Word order: “the kitchen light” can be changed to “the light in the kitchen"
- Singular/plural: “the kitchen light” can be changed to “the kitchen lights”
- Questions: “close the door” can be changed to “do you mind closing the door?”
- Negation: “turn on the tv at 19:00” can be changed to “don’t turn on the tv at 19:00”
- Politeness: “turn on the tv” can be changed to “could you please be so kind as to turn on the tv?”

The user can ask to the NLG process to generate only some of these variants: for example, just “word order” and “questions”, but no “negation”, “singular/plural” changes nor “politeness”.

Also, each generated sentence is assigned a number that expresses its “complexity” level, that is, how much has been modified with respect to the original one. The user can ask to NLG process to generate only the least complex sentences (10-20) or the whole range of them (which can be hundreds of sentences), depending on their needs or their chatbots’ capacity.

The result is a powerful but flexible system that can be configured to generate as many, or as few, variants of a sentence as the user needs.