Data prerequisites

One of the important benefits of the Pirkin models is the fact that they work with limited datasets/corpuses. This doesn’t mean that we don’t need important data, but the required data already exists:

1. for all the languages (as they are prerequisites in the learning curricula of any school),
2. at the company level (especially that we give alternatives),
3. or, it can be easily created as the models are working with sentence-intents that is simplifying the structuring/labeling of the data.

**NIU-NLU Specific** **datasets** (main tasks, specific for the bot; one of the following or a combination of them):

* Corpus of an existing site (or sites for multi-domains bot)
* Unlabeled logs from the current relations with the clients
* CRM data
* Books, documents related to the main tasks of the bot.
* Existing corpuses related to the task/core-input

These datasets are important for the language model, extraction of the answers for back-up models, extraction of the core-vocabulary and extraction of the named entities. If none of these are available, the vocabularies need to be provided independently and back-up models will not be effective. Even so, the Pirkin models can be implemented.

**NOG-NLG Labeled datasets** (specific for the bot main tasks; one of the following or a combination of them):

* Q&A labeled data. Existing or created by screenwriters (the question can be sentence-intent type: 1 Subject + 1 Predicate + 1 Complement + 1 Attribute.
* Chitchats labeled data. Existing or created by screenwriters split in 8 categories[[1]](#footnote-1) + feedback (with 4 sub-categories) + elaboration database that is specific for queries (with 7 subcategories)– see Labyrinth model.
* Reactions/Replies labeled data. Existing or created by screenwriters. Needs to cover 3 types of replies: confirmations / satisfactions/ understanding.
* Specific answers for expressions/ironies…

For chitchats or replies the model permits data augmentation methods from other languages.

**General vocabularies**:

* The extended vocabulary of a language with all the words and their POS
* Synonyms/neologisms/regionalisms vocabularies
* Connectors (ex: in, at, the, will…)
* All verbs with all conjugations/forms for (future/present/past, first person/second/third, affirmative/negative forms), or effective solutions, depending on the language, to be extracted from infinitive form.
* Pronouns with all forms
* Dictionaries for expressions, adverbs and superlatives
* Emoji dictionaries
* Vocabularies for negations/affirmations/questions
* Dictionaries for numerical data and special dates/events (ex: Monday, January, Eastern…)

These vocabularies can be extracted if there is a comprehensive corpus for the language or a pre-trained model available.

**Specific vocabularies:**

* Extracted from the NIU-NLU Specific datasets or provided
* Specific adjectives (related with the bot task) – can be extracted
* Abbreviations dictionaries.
* NER- entities vocabularies, extracted or provided.
* GER- generalities database. They are linked with the bot task, but are considered generalities topics
* Prioritization of NER in terms of the bot tasks. For example, in the financial advisory domain, bond and equities will have their own classes and we need to establish if the brokerage firms are entities more important for our tasks or the name of the equities. The NER architecture/prioritization needs to be decided depending on the bot tasks (especially if we are dealing with a multi-domain bot in which domains can have inter-dependencies).

**Deep conversational datasets:**

* Books related to the bot task (for example a gastronomy book for restaurant recommendation bot)
* Or books to be used as out-of-topic existential/additional discussions that can benefit the conversation.
* Extract Books titles/ chapter titles/dialogues/first phrase from each chapter/sub-chapter

If effective models have already been created for a specific language or domain, using Cezanne-ai open-framework, then most of these prerequisites are already available (without further research), if open-sourced. Oscar or Common Crawl can be also used for existing corpuses.

**Bot random questions/disclaimers/change topics:**

The model is working with pre-defined generated inputs from the bot (see specific pipelines in the Cezanne-ai project layer: NOG/NLG). For that reason, we need to provide at least 5 choices for each category in the needed language and adapted to our business/PR/marketing objectives.

1. See Diatribe policies [↑](#footnote-ref-1)