**Description de l’approche :**

Pour annoter notre corpus en actes de dialogues, nous avons fait le choix d’utiliser des règles heuristiques. Cette méthode permet d’une part d’éviter l’annotation manuelle couteuse et laborieuse, mais surtout d’adapter l’annotation à d’autres contextes, et d’obtenir instantanément les labels de nouvelles données si le corpus est élargi par la suite.

Cette flexibilité et automatisation de l’annotation du corpus nous est permise par Snorkel. Cet outil facilite la création de règles à partir de données très disparates, puis crée un modèle génératif qui va traiter les labels bruités attribués par les différentes règles. En sortie du modèle, on obtient la distribution de probabilités des labels pour chaque élément de notre corpus. L’étape finale consiste à donner les données annotées ainsi obtenues à l’entrainement d’un algorithme de machine learning plus classique, qui ne nécessiterait par exemple dans notre cas que de la transcription textuelle.

1. Écriture des règles

Pour écrire les *labeling functions* (LFs) qui correspondent à nos règles heuristiques, nous avons observé uniquement notre set de développement constitué de x% de notre corpus. Contrairement au langage écrit, les transcriptions de langage oral spontané ne

1. Modèle génératif