1. Entities (The main words of our natural language can be divided into following categories. Most words which are not is these categories would be labeled with “NULL”)
   * Dimensions (BusinessPartner, Item, etc.)
   * Measures (SalesAmount, Balance, etc.)
   * DimensionValue (placeholder for DB scanned value, e.g. #itemcode1)
   * Numbers (1, 3, one, three, etc.)
   * Datetime (date, time, period etc.)
   * GeneralOperators (>, <, =, !=)
   * DatetimeOperators (before, in, after, from…to, until, etc)
   * Aggregations (Sum, average, max, min, count, etc.)
   * Orders (Top, Bottom, most, least, best, worst, etc.)
2. Tags (We defined the following tags for our NER model. All of them are used to translate the natural language to an SQL qury)
   * B-Field, I-Field
     + Dimensions
       - List all prospects (B-Field)
     + Measures
       - Procurement price (B-Field, I-Field)
       - What is the purchase price of product N (B-Field, I-Field)
   * B-GroupBy, I-GroupBy
     + Dimensions
       - What is the first deal today (B-GroupBy)
       - List the top 3 sales person (B-GroupBy, I-GroupBy)
   * B-Order, I-Order
     + Orders
       - List the top 3 sales person (B-Order)
       - Which product has the biggest back order quantity (B-Order)
   * B-OrderBy, I-OrderBy
     + Dimensions
     + Measures
       - Which product has the biggest back-order quantity (B-OrderBy, I-OrderBy)
       - List the top 3 vendors by amount (B-OrderBy)
   * B-Limit, I-Limit
     + Numbers
       - List the top 3 sales person (B-Limit)
   * B-Aggregation, I-Aggregation
     + Aggregations
       - What is the total back-order quantity (B-Aggregation)
       - What is the average purchase price of product N (B-Aggregation)
   * B-GeneralFilterAggregation, I-GeneralFilterAggregation
     + For filter on measures,
       - List the items whose revenue are larger than 1 million (GeneralFilterAggregation is the default “sum”)
   * B-GeneralFilterField, I-GeneralFilterField
     + Dimensions
       - What is the average purchase price of product N (B-GeneralFilterField)
     + Measures
   * B-GeneralFilterValue, I-GeneralFilterValue
     + Value of ‘GeneralFilterField’, string or DimensionValue
       - What is the average purchase price of product N (B- GeneralFilterValue)
       - What is the sales amount of BP\_NAME\_CUSTOMER (B-GeneralFilterValue)
       - What is the sales amount of ITEM\_NAME (B-Gene
       - ralFilterValue)
   * B-GeneralFilterOp, I-GeneralFilterOp
     + GeneralOperators（No examples in current corpus.）
   * B-GeneralFilterAggregation, I-GeneralFilterAggregation
     + Aggregations(No examples in current corpus.)
   * B-DateTimeFilterValue, I-DateTimeFilterValue
     + Datetime
       - What is the total purchase quantity of product A this month (B-DateTimeFilterValue, I-DateTimeFilterValue)
   * B-DateTimeFilterField, I-DateTimeFilterField
     + The qualifier of Datetime
       - Show balance of B/S account X in fiscal year Y (B-DateTimeFilterField, I-DateTimeFilterField)
   * B-DateTimeFilterOperator, I-DateTimeFilterOperator
     + DatetimeOperators
       - Show the sales amount from 2015 to 2017 (B-DateTimeFilterOperator, B-DateTimeFilterOperator)
       - Show the sales amount till last year (B-DateTimeFilterOperator)
   * B-ComparisonWith, I-ComparisonWith
     + Compare the total purchase quantity of product A with product B (B-ComparisonWith, I-ComparisonWith)
     + Compare balance of B/S account X with last year (B-ComparisonWith, I-ComparisonWith)
3. Entire Examples (with format)
   * Top 3 sales order.
     + Top B-Order
     + 3 B-Limit
     + sales B-GroupBy
     + order I-GroupBy
   * Top 1 vendors by quantity.
     + Top B-Order
     + 1 B-Limit
     + vendors B-GroupBy
     + by NULL
     + quantity B-OrderBy
     + Preposition here means nothing, labeled with ‘NULL’
   * How many employees.
     + How B-Aggregation
     + many I-Aggregation
     + employees NULL
     + “Employees” here could be labeled as B-Field, but “Employees” is not a specific field of database view, so we label it with ‘NULL
   * What is the average purchase price of product N buyer A purchased.
     + What NULL
     + is NULL
     + the NULL
     + average B-Aggregation
     + purchase B-Field
     + price I-Field
     + of NULL
     + product B-GeneralFilterField
     + N B-GeneralFilterValue
     + buyer B-GeneralFilterField
     + A B-GeneralFilterValue
     + purchased NULL
     + Words like “what”, “is” and “the” should be labeled as ‘NULL’
   * Which product has the biggest back-order quantity.
     + Which NULL
     + product B-GroupBy
     + has NULL
     + the NULL
     + biggest B-Order
     + back-order B-OrderBy
     + quantity I-OrderBy
   * What is the GRPO amount of this month.
     + What NULL
     + is NULL
     + the NULL
     + GRPO B-Field
     + amount I-Field
     + of NULL
     + this B-GeneralFilterField
     + month B-GeneralFilterValue
   * The average purchase price of product A vs. the same period last year
     + The NULL
     + average B-Aggregation
     + purchase B-Field
     + price I-Field
     + of NULL
     + product B-GeneralFilterField
     + A B-GeneralFilterValue
     + vs. NULL
     + the NULL
     + same B-ComparisonWith
     + period I-ComparisonWith
     + last I-ComparisonWith
     + year I-ComparisonWith
4. Instruction
   * First, try to translate your natural language to SQL query, then the item you want to select is “B-Field”. If there is an “Order” in your query, then the item might be labeled with “B-GroupBy”. In the same way, the words following “where” should be labeled with “B-GeneralFilterField” or “B-GeneralFilterValue”
   * Second, words like “Which”, “What”, “is”, “the” and meaningless preposition in you SQL query, should be labeled with “NULL”, if there are punctuation mark in your sentences, label them with “NULL” as well
   * Third, the adjective of “B-Field” should be labeled with “NULL” if it’s not included in database field
   * Forth, The format of your corpus should be same with above entire examples. Each line has one word and its label splited with a single blank space. There should be a space line between to sentences.
   * Fifth, “B-“ means begin of an entity, “I-“ means the following words belong to that entity.
5. Limition
   * Only support question with one aggregation method, though measures can be multiple
   * Do not support “trend”, for it’s a rare case
   * Only support one order by
   * Only support one date time filter
   * Only support one group of comparison
     + What’s the total sales amount of iphone this year comparing with last year, and comparing SangSum (Be not supported)
6. When a sentence come in
   * Preprocessing
   * Tokenization
   * Topic Found
   * NER(Name Entity Recognition)
     + If topic have default values such as default fields, default aggregation method or default filter, using the default as final entities. Otherwise, using the NER results.
   * Intent Format (Find default value according to topic)
     + Intent Name
     + View
     + Default Dimension
     + Default Measure (appear with Aggregation)
     + Default Aggregation
     + Default Filter (Option)