lang.ai

Complex customer interactions into effortless structured data.

Technology

WHITE PAPER

Pre-processing

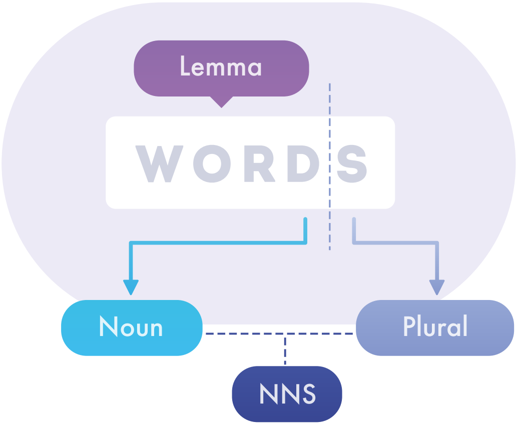
The process where we prepare the data to be ingested by the algorithm. Both tokenization and automatic correction are language-agnostic, whereas *lemmatization is only available in certain languages*. Our algorithm can work only with tokens without the information being lemmatized.

01 Tokenization

Tokenization is a preprocessing method that is aimed at extracting the most simple features that compose a text. Tokenizing means splitting your text into minimal lexicographic units, and it is mandatory before any kind of text processing.

02 Lemmatization

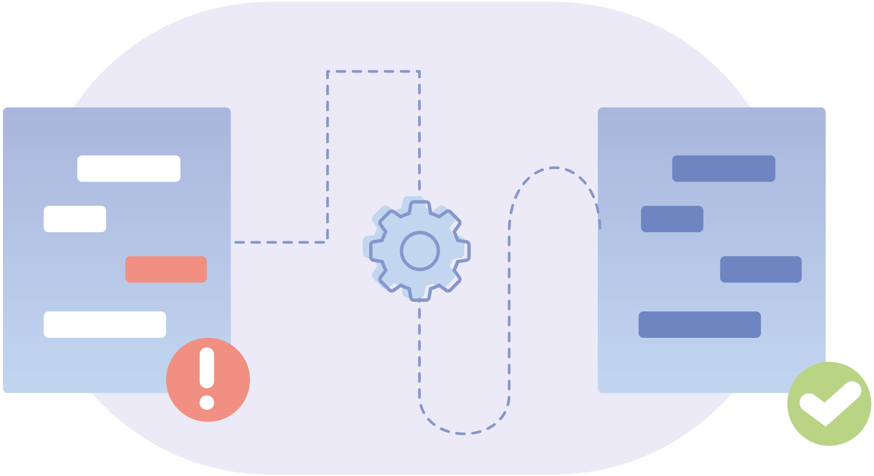
Lemmatization is a normalization operation that has to do with transforming the sequence of words composing a text by removing the inflectional ending



of each word (usually, by combining a morphological analysis with PoS-tagging). The result is a new sequence composed of basic (or dictionary) word forms that are also called word lemmas.

03
Automatic correction

We create statistical language models of word pairs based on the distributions of words and their occurrence contexts. From such models, we establish a series of conditions based on information theory for evaluating, detecting and replacing possible misspelled words. Two different approaches are regarded to address these operations: a probabilistic weighted distance approach and a RNN-based approach.



Discovery

The process where we execute our algorithm to extract the main intents and features in a dataset. It is mostly based on Information Theory.

04

Intent induction

It is our unsupervised categorization process that takes as input a raw collection of text documents and produces a set of text categories called intents. Each intent is represented by a set of lexically related words in the context of the input collection.

To induce intents:

- It first calculates a set of semantic edges where each edge is a pair of words deemed to entail more information than either the first or the second word by separate.

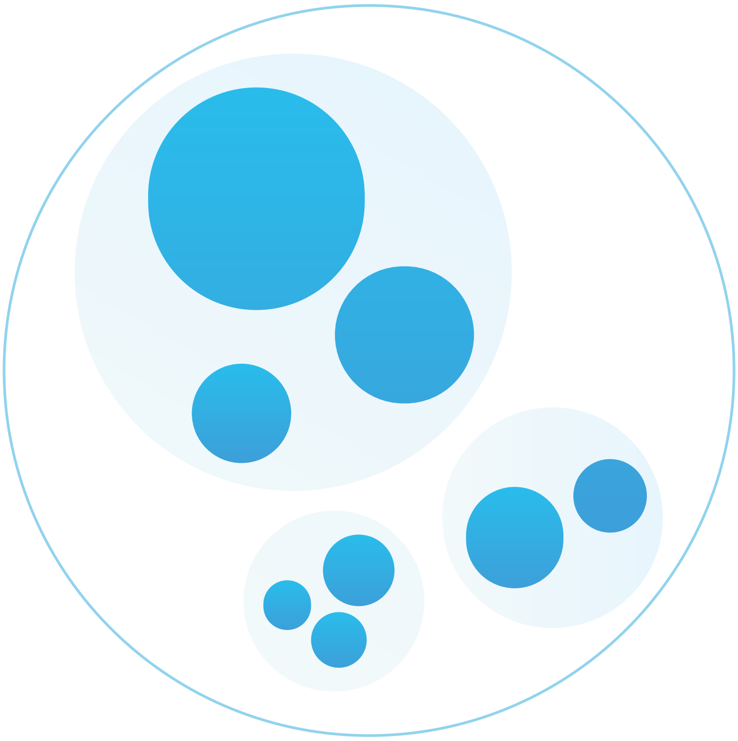
- Then, signatures are defined as cliques of words extracted from the input documents via the semantic edge relationship.

- Finally, signatures are organized in a generalization DAG and those more general ones are regarded as intents.

05

Semantic relation and clustering

A clustering process is performed over the discovered intents in order to provide the user with a higher level categorization. Thus, intents that share a common (underlying, but interpretable) concept are grouped together in a new named category whose label (i.e., a set of words) describes the relationship among the grouped intents.



For example, in a corpus containing tweets about an air carrier, we can find different intents that are somehow related to flight delays, usually due to air traffic conditions: “delay-hour“, “miss-connection“ and “air-traffic“. By applying this semantic clustering, we are able to group the three intents in the same cluster, and also determine that the word “delay“ plus the name of the air carrier will form the label with which this higher level category will be named.

06

Features extraction and deeper intents

Intents also might consist of features; i.e., words that individually can be used to disambiguate or provide a more specific interpretation of the intent in the context of the input corpus.

Such features are inferred in an unsupervised way by using a similar approach to that of intent induction (regarding in this case those texts in the coverage of the intent).

Postprocessing

The process where we add information into the algorithm once the discovery phase has finished. It may be a one-time or a recurring process

07

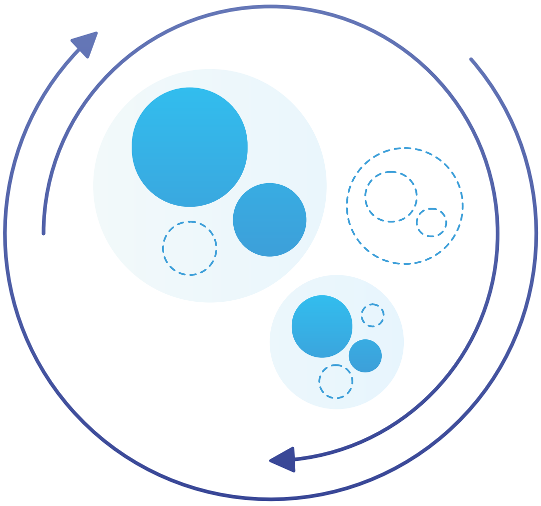
Augmented knowledge

With the aim of providing our users with a more accurate and robust classification models, we perform a knowledge augmentation operation by enriching each intent model with external information from a general-purpose knowledge base, namely, Wikipedia.

To enrich each intent, we firstly select those concepts from the external source that are closer to the original intents of a corpus. In this process we perform an exhaustive analysis of the intents in order to retrieve only those concepts that are highly related to each intent, trying to eliminate all the possible ambiguities that can be introduced. Then, we enrich each intent by combining the corpus-based model with a model of the selected concepts from the external source.

08 Adaptive learning

We fit our intent induction approach into an adaptive learning framework that allows the inference of text categories incrementally over time. This allows us to discover and improve the current intents when the user uploads new datasets. One example may be to upload data from a different quarter.



We provide an adaptive algorithm whose intents model is a mixture between the previous intents model and the intents model extracted from the current corpus. The mixing coefficient can be fine tuned to adjust how much weight we want to assign to historic data versus the new data.

As a result, the algorithm is able to detect new intents and transform or augment previous intents with information extracted from the new dataset.

09 Real-time classification API

To facilitate the use of the inferred intents, we built a set of intent classifiers that allows to label new (incoming) texts with intents. These classifiers are exposed behind an API where users can make secure requests to classify new messages. The user may also remove, edit or create new rules to fine-tune the classifier.

lang.ai

Complex customer interactions into effortless structured data.

SMART
We understand and tag your customer interactions, only from your own private data.

CUSTOMIZABLE
You know your business to decide what to measure.

SCALABLE
We make scalability easy so your CS team stays focused on your customers.

EFFORTLESS
Instant connection, results in minutes, no need of AI / coding knowledge.