ACTIVE ANNOTATION

BOOTSTRAPPING ANNOTATION LEXICON AND GUIDELINES FOR SUPERVISED NLU LEARNING

Federico Marinelli, Alessandra Cervone, Giuliano Tortoreto, Evgeny A. Stepanov, Giuseppe Di Fabbrizio, Giuseppe Riccardi

VUI Inc, United States — SiS Lab, University of Trento, Italy

Background & Motivations

Research Questions

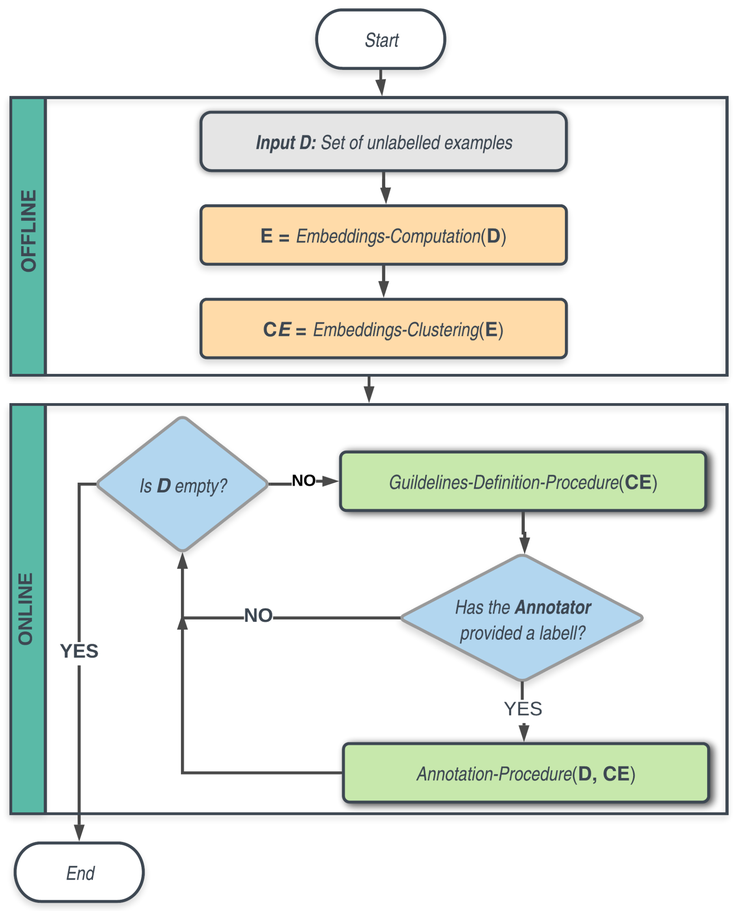
- Can we learn annotation lexicon and guidelines directly from data?
- Can we improve manual annotation accuracy and increase annotation speed by leveraging ML techniques?
- What is the optimal balance between accuracy and speed in human annotation tasks?

Proposed Solution

Learn the *y*-label space on-the-fly during the annotation process by combining human and machine intelligence:

- Unsupervised learning in the embedding space
- Human-in-the-loop verification process
- Linguistic insights to create lexicons that can be open categories and adapted over time

Active Annotation Algorithm Overview



Guidelines Definition Procedure

Exploration Phase

Represents the process by which the annotator, given *N* Pivot data points and an automatically computed cluster label, is asked to provide a *y*-label to the given *N* Pivot data points.

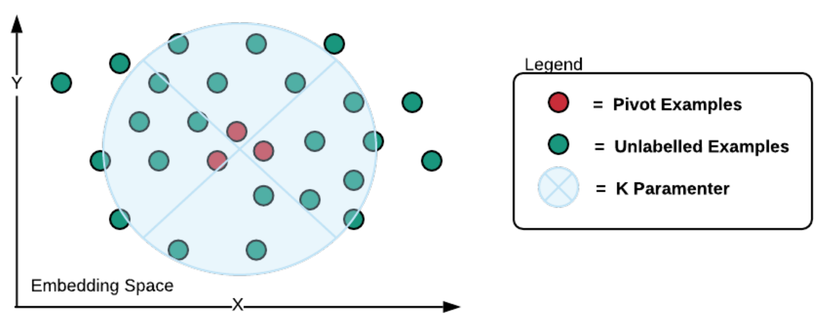
Algorithm 1: *Guidelines Definition Algorithm Overview*

Input = $C_{E'}$ – Clustered input data;
  Select one random cluster c inside $C_{E'}$;
  Select the N most informative data points in c as Pivots;
  Compute a cluster label L_c;
  Propose the N Pivots data points and L_c to the annotator;
if *Annotator provides a label* **then**
  return True;
end

Annotation Procedure

Exploitation Phase

- In case the annotator provides a *y*-label, the algorithm proposes *K* data points, that are very similar to the *N* Pivot data points.
- The annotator is asked to analyze the proposed *K* data points and annotate them through a binary decision process.



Experimental Design

Baseline

The annotator is asked to label one randomly selected sentence at a time.

Dataset

Sub-sample of Microsoft e2e Dialogue Challange Dataset — movie domain.
Statistics: 2k sentences — 14 Ground-Truth Intent Labels

Experiments

Four experiments with four Annotators:
Phase 1: Annotation using the Active Annotation Web-Tool (25 minutes)
Phase 2: Annotation using the Baseline Annotation Web-Tool (25 minutes)

Results

	Baseline		AA	
	μ	σ	μ	σ
Number of sentences labelled in 25 mins	118.6	18.5	999.3	171.4
Number of distinct labels	10.3	1.2	8.6	0.9
10-Fold Cross-Validation F1	0.83	0.05	0.91	0.02

	Baseline		AA	
	μ	σ	μ	σ
F1-score Test-Set	0.81	0.04	0.89	0.03
Annotation Agreement	0.61	0.02	0.64	0.01