Leveraging Sequential Pattern Information for Active Learning from Sequential Data

Raúl Fidalgo-Merino, European Commission (JRC)

ICPR 2020 (Online), 15/1/2021

Joint
Research
Centre

Contents

- Introduction
- Active Learning (AL) techniques for non-sequential and sequential data
- SPIAL: A new AL approach for training sequential models
- Experiments
- Conclusions and future work

Introduction

- Evolution of technologies → Production of huge amounts of sequential data
- Data analysis techniques can extract information from these data (e.g., supervised machine learning methods for sequential data)
- Some of these techniques need proper training data to construct models accurately
- Active Learning techniques for sequences selects training data:

↓ annotation costs, ↑ model performance

Active Learning techniques

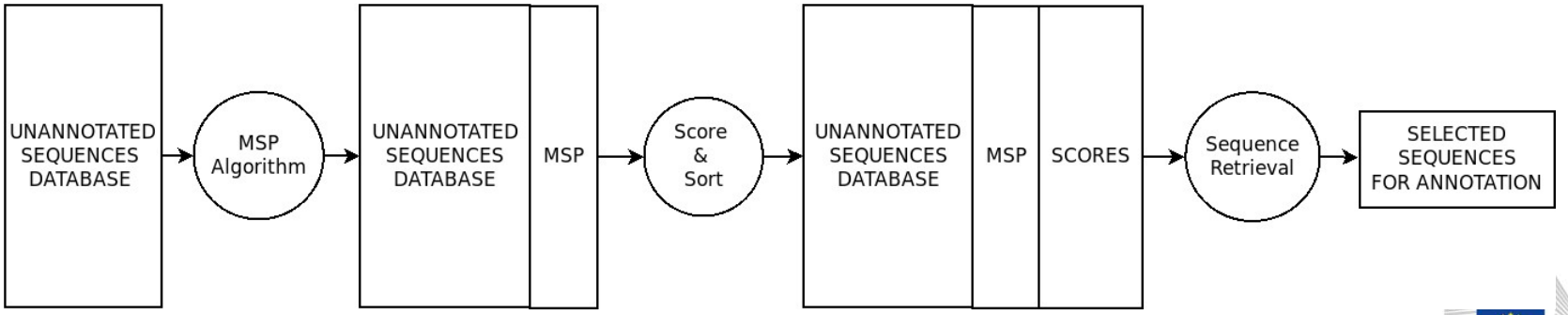
- Active Learning (AL) techniques usually based on three principles:
  - Representativeness: Data selected for training to cover representative sections of the instance space
  - Diversity: Data selected as dissimilar as possible to avoid overfitting
  - Uncertainty: Data selected iteratively based on internal metrics of a model
- Some techniques cover more than one principle → Hybrid techniques
- Drawbacks: High computational cost (repr. and div.), slow learning curve and/or overfitting (unc.)

Active Learning for sequential data

- These principles are also applied to AL techniques dealing with sequential data
  - Representativeness- or diversity-based techniques usually employ Cosine Similarity with DTW to decide when two sequences are similar
    - High computational cost → Data pool, feature independent
  - Uncertainty-based techniques uses a base learner to select training data based on its performance (iterative strategy)
    - Slow learning curve, tend to overfit
- Hybrid techniques produce more robust and better selection strategies but can inherit several drawbacks

SPIAL: A new AL approach for training sequential models

- Proposed technique: SPIAL (Sequential Pattern Information for Active Learning)
  - Uses Sequential Pattern Mining techniques
  - Based on representativeness and diversity principles
  - Modular and less computational cost than other repr.- and div.-based techniques



SPIAL (step 1): Extraction of sequential patterns

- Uses Sequential Pattern mining techniques to extract representative information from the sequential data base

$$sp\ is\ Sequential\ Pattern\ \leftrightarrow\ |\{s\}|\ \geq min\_sup, s \in S, sp \subseteq s$$

  - *S*: sequential database, *min_sup* is a parameter to decide when a sub-sequence is relevant

  - Particularly, SPIAL extracts Maximal Sequential Patterns

$$sp\ is\ maximal \leftrightarrow\ \nexists sp' \in S / sp \subseteq sp'$$

SPIAL (step 2): Scoring sequences by representativeness

- Each sequence in the database of sequences is scored based on the sequential patterns found in them. Examples:
  - *CountSP* criterion: A sequence obtains a score equals to the number of maximal sequential patterns that it contains

$$CountMSP(s) = |\{msp_i\}| / \forall msp_i \in MSP, s \subseteq S$$

  - *MSPCoverage* criterion: A sequence is scored based on the coverage of the sequential patterns that it contains

$$MSPCoverage(s) = \left|\{s_j\}\right| / \forall msp_i \in MSP, \forall s_j \in S, msp_i \subseteq s \wedge msp_i \subseteq s_j$$

where: *MSP* is the set of Maximal Sequential Patterns extracted and *S* is the sequential database

SPIAL (step 3): Retrieving diverse sequences

- After sorting the scored sequences descending order, they are selected according to a exclusiveness criterion
  - A sequence is selected for annotation if its maximal sequential patterns did not appear in a previous selected sequence
  - A sequence cannot be selected more than once
- If all maximal sequential patterns are contained in the selected sequences but there is more room for additional sequences for annotation, SPIAL empties the set of covered maximal sequential patterns and continues with the selection procedure from the beginning of the list
- The final output of SPIAL is a list of proposed sequences for annotation naturally sorted by their representativeness and diversity

Experiments (I)

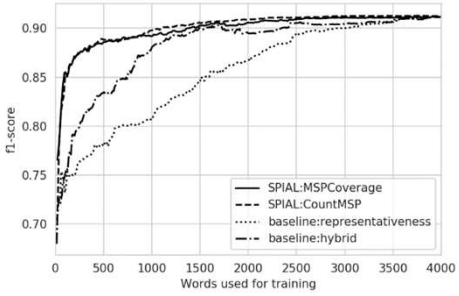
- Three datasets: Logistic transportation, CoNLL2002 and CoNLL2003
- Three Active Learning techniques for sequential data under evaluation:
  - SPIAL (2 versions)
    - *VMSP* algorithm for Maximal Sequential Pattern mining
    - Scoring functions: *CountMSP* and *MSPCoverage*
  - Representativeness-based: using cosine-similarity and DTW
  - Hybrid (repr. and diversity): cosine similarity and DTW with similarity threshold (β)
- Base Machine Learning algorithm for sequential data: CRF (considering 3 events previous and next to the current one)

Experiments (II)

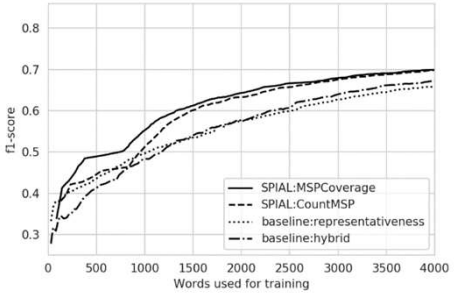
- 10-fold cross-validation
- Parameter setup using a Grid Search technique
  - Logistic transportation: min_sup=50 (SPIAL); β=0.9 (representativeness and hybrid)
  - CoNLL2002: min_sup=25 (SPIAL); β=0.9, Pool=4K (representativeness and hybrid)
  - CoNLL2003: min_sup=25; β=0.9, Pool=4K (representativeness and hybrid)
- Evaluation metrics
  - Model assessment using selected sequences: *f1-score*
  - Execution time of selection strategies

Experiments (I): Performance

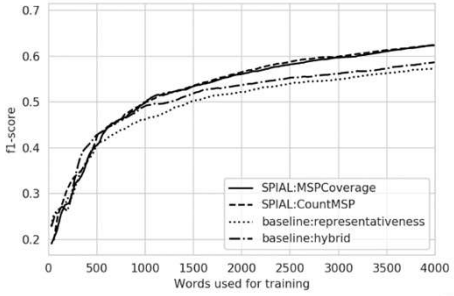
- Significantly better convergence speed of both versions of SPIAL
  - Needs half of the sequences to reach the performance of representative-based and hybrid techniques
- Models trained using sequences selected by SPIAL shows a better convergence speed as well as a higher performance than those using the other AL techniques evaluated



Logistic transportation



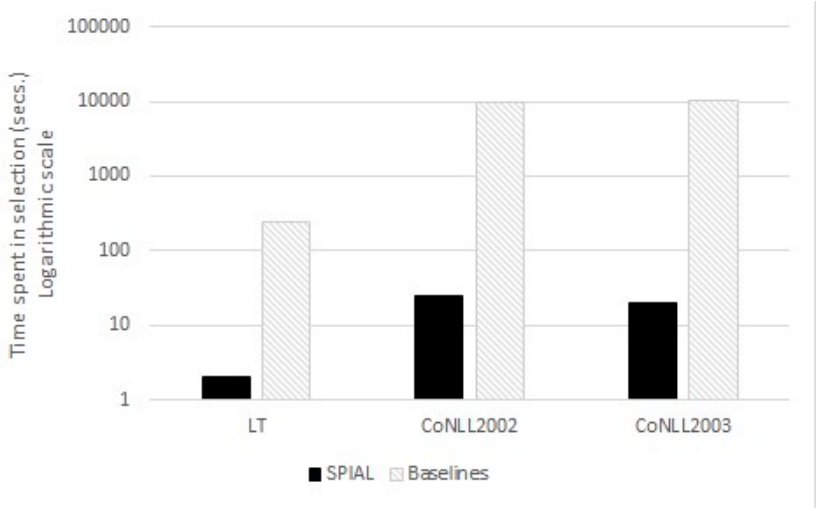
CoNLL2002



CoNLL2003

Experiments (II): Execution times

- Execution times of SPIAL in the three experiments reveal important savings compared to state-of-the-art (soa) techniques
  - Logistic Transportation: SPIAL is two orders of magnitude better than soa
  - CoNLL2002 and CoNLL2003: SPIAL is three orders of magnitude faster than soa



European Commission

Conclusions and future work

- SPIAL is a novel Active Learning technique for sequential data
  - Based on representativeness and diverse principles
  - Uses Sequential Pattern information contained in the database of sequences
  - It is feature independent and its modular design allows extensions of the methodology
- According to the experiments performed, SPIAL showed that:
  - Its execution times were between 2 and 3 orders of magnitude less than soa techniques
  - Machine learning models trained have better convergence speed and performance in general than soa techniques evaluated

Thank you

European Commission