Leveraging Sequential Pattern Information for Active Learning from Sequential Data

R. Fidalgo-Merino, L. Gabrielli, E. Checchi

Motivation

Thanks to the recent evolution of technologies, IT systems are now capable of generating high amounts of sequential data. These data contain valuable information that can be extracted by supervised machine learning techniques for sequential data.

Active learning techniques for sequential data are needed to select unlabeled sequences for annotation and training of machine learning models, reducing human interaction and improving its performance.

Current state-of-the-art active learning techniques for sequential data perform better than random selection but they have several drawbacks like, high computational cost, feature dependency (require to know the features taken into account by the learner) or slow learning curve.

SPIAL: A new AL approach for training sequential models

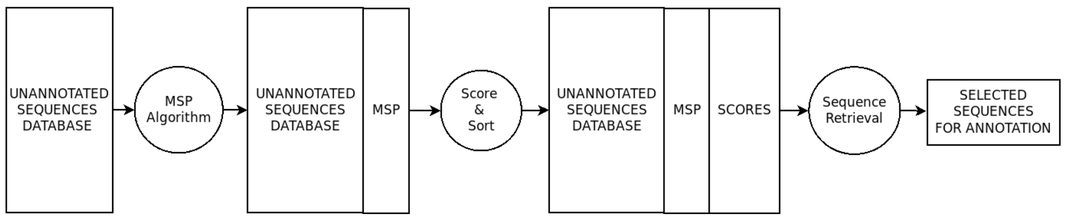
We present a novel active learning technique for sequential data based on sequential pattern information extracted from the available database of unlabeled sequences.

SPIAL (Sequential Pattern Information for Active Learning) is aimed at obtaining accurate machine learning models:

- with fast convergence speed,
- reduced computation time and,
- feature independency

The proposed technique can be summarised in three steps:

1. Extraction of maximal sequential patterns (MSP) from the database of unlabeled sequences
2. Scoring each sequence in the database based on the representative information contained in them, for instance
   1. *CountMSP*: number of MSP in each sequence
   2. *MSPCoverage*: MSP coverage present in each sequence
3. Retrieval of sequences with diverse MSP



The final output of SPIAL is a list of proposed sequences for annotation naturally sorted by their representativeness and diversity.

Empirical validation

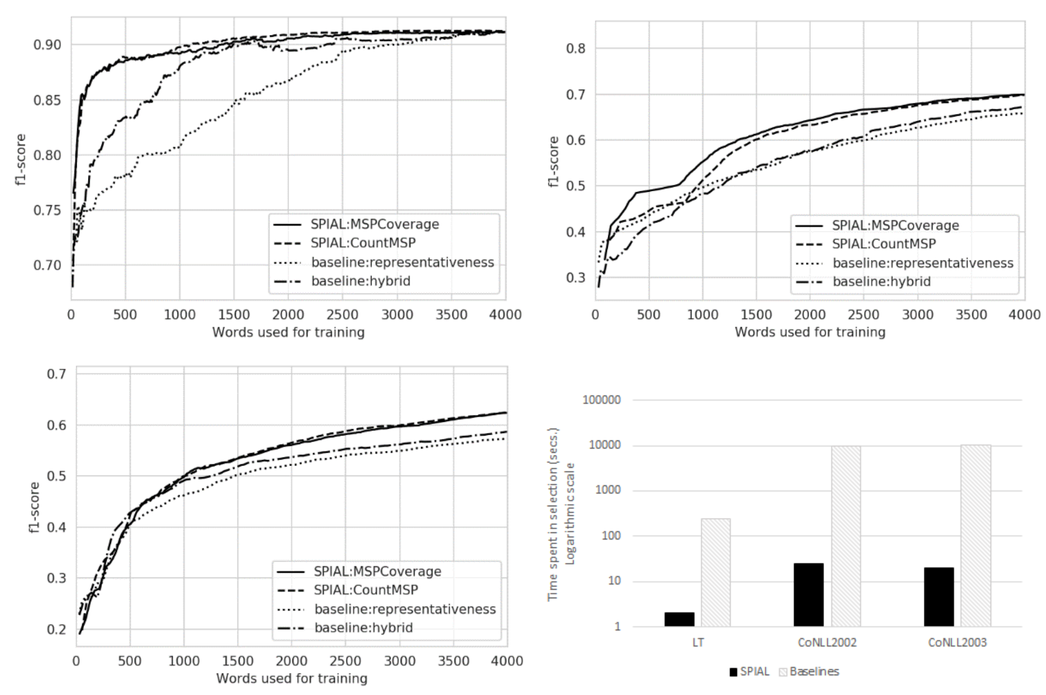
SPIAL has been assessed using three different datasets:

- Logistic transportation
- CoNLL2002
- CoNLL2003

and two State-Of-the-Art (SOA) active learning techniques for sequential data based on extracting the representativeness and diversity of sequences using cosine similarity.

The base learner trained with the sequences selected by the different active learning methods under evaluation was Conditional Random Fields (CRF).

The experiments followed a 10-fold cross-validation and the evaluation metrics chosen were f1-score and execution time. The next figure shows the results on performance (logistic transportation – top left; CoNLL2002 – top right; CoNLL2003 – bottom left) and execution times (bottom right) obtained.



These results reveal that SPIAL outperforms the performance of current SOA techniques in the field but also that its execution times are several order of magnitude faster.

Conclusions

This paper presents SPIAL, a novel technique for Active Learning from sequential data based on Sequential Pattern information contained in the database of unlabeled sequences available. It is feature independent and its modular design allows to extend its functionalities easily. According to the experimental results, SPIAL selects faster the sequences to train than current SOA techniques and they produce models with better performance.

The European Commission's science and knowledge service
Joint Research Centre

EU Science Hub: *ec.europa.eu/jrc* @EU_ScienceHub EU Science Hub
EU Science Hub - Joint Research Centre EU Science, Research and Innovation

Joint Research Centre

Contact:
Raul Fidalgo-Merino (raul.fidalgo-merino@ec.europa.eu)
Scientific/Technical Project Officer
European Commission, Joint Research Centre (JRC), Ispra, Italy
Text and Data mining Unit, Competences Directorate