A Greek Named-Entity Recognizer that Uses Support Vector Machines and Active Learning

Georgios Lucarelli and Ion Androutsopoulos

Dept. of Informatics, Athens University of Economics and Business
Patision 76, GR-104 34, Athens, Greece

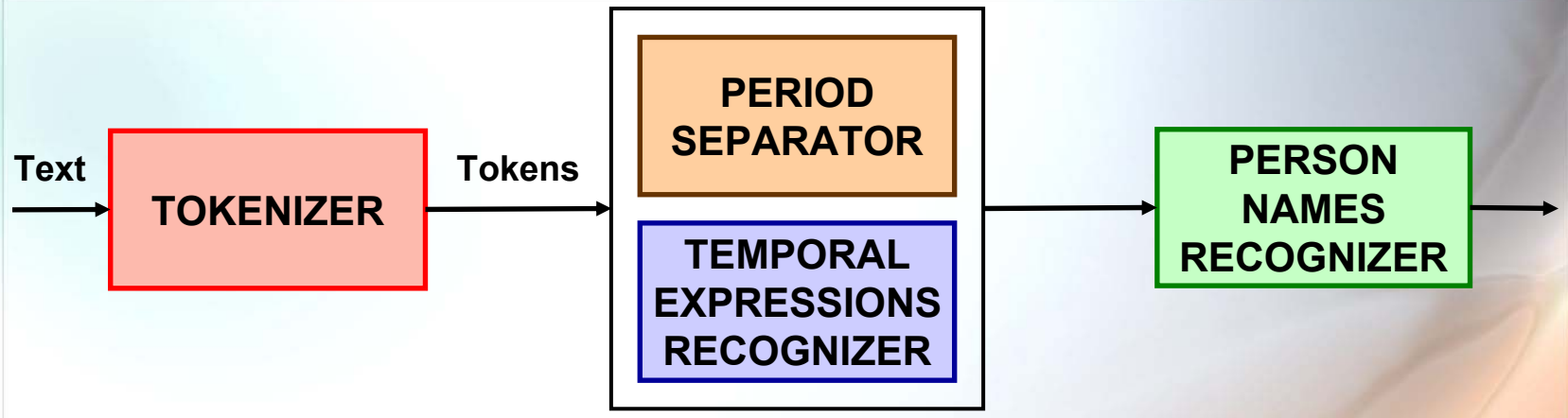
Introduction

- **The problem:**
  - **Find and categorize in predefined categories Greek named entities appearing in documents.**

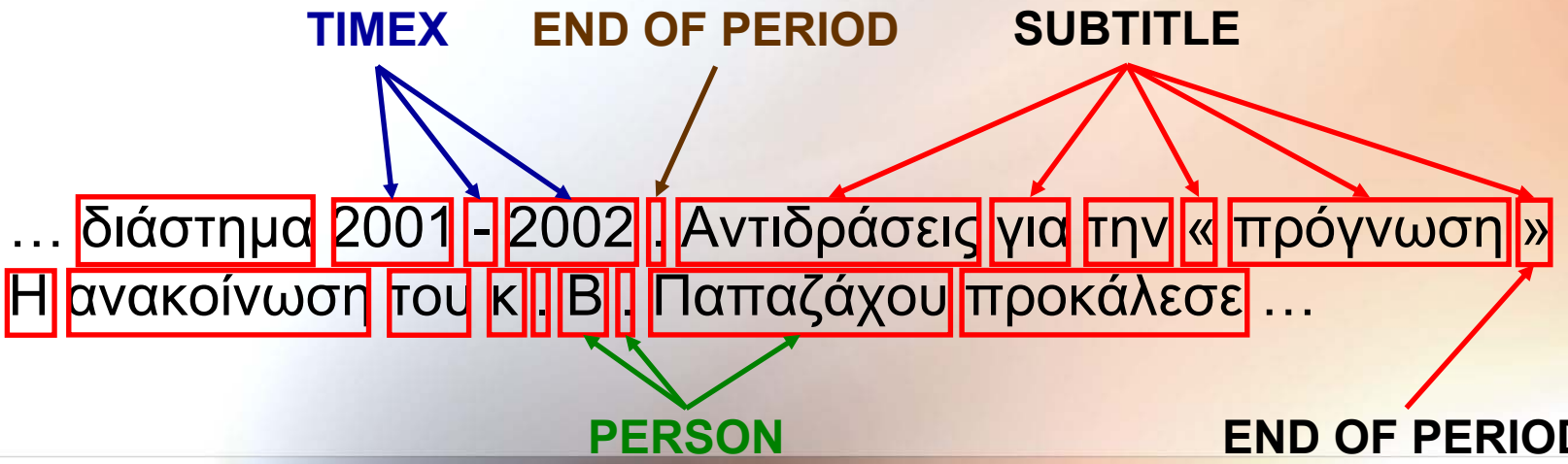
Τον <TIMEX TYPE="DATE">Απρίλιο του 1948</TIMEX>, ο <ENAMEX TYPE="PERSON">Πολκ</ENAMEX> έρχεται για δεύτερη φορά στην <ENAMEX TYPE="LOCATION">Ελλάδα</ENAMEX> και παντρεύεται την αεροσυνοδό της <ENAMEX TYPE="ORGANIZATION">ΤΑΕ</ENAMEX> (της μετέπειτα <ENAMEX TYPE="ORGANIZATION">Ολυμπιακής Αεροπορίας</ENAMEX>) <ENAMEX TYPE="PERSON">Ρέα Κοκκώνη</ENAMEX>, που είχε γνωρίσει στο προηγούμενο ταξίδι του.

- Different approaches:
  - systems with hand-crafted rules
  - machine learning systems
  - hybrid systems.

System Architecture

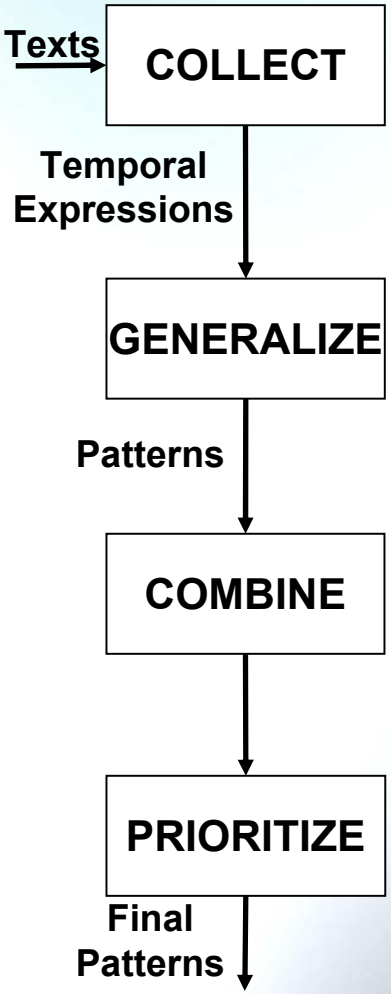


… διάστημα 2001-2002. <SUBTITLE> Αντιδράσεις για την «πρόγνωση» </SUBTITLE> <PARAGRAPH> Η ανακοίνωση του κ. Β. Παπαζάχου προκάλεσε …



Temporal Expression Recognition

- **Semi-automatically produced patterns**



Τετάρτη 31 Ιανουαρίου 2001

1/12/2005

10.1.2005

DAY [0-9]{2} MONTH [0-9]{4}

[0-9]{1} SEPARATOR [0-9]{2} SEPARATOR [0-9]{4}

[0-9]{2} SEPARATOR [0-9]{1} SEPARATOR [0-9]{4}

([0-9]{1}|[0-9]{2}) SEPARATOR ([0-9]{1}|[0-9]{2}) SEPARATOR [0-9]{4}

… η συνέλευση θα γίνει στις 31 Ιανουαρίου 2001 …

[0-9]{2} MONTH [0-9]{4}

[0-9]{2} MONTH

[0-9]{4}

Imbalance of Categories

- The ratio of person-name tokens to non-person name tokens is 1:42.
- Classifiers may learn to classify always in the most frequent category.
- Using “sure-fire rules” to identify non-person names, the ratio becomes 1:3.5.
- Example:

… Όπως αποκάλυψε ο κ. Βασίλης Παπαζάχος, στις 14 Σεπτεμβρίου 2000 η επιστημονική του ομάδα έστειλε άρθρο στο περιοδικό "Geophysical Journal International", ενημερώνοντας για σεισμική δόνηση με επίκεντρο την Τάφρο του Βορείου Αιγαίου …

Domain-independent Features

t_{-2} t_{-1} t_0 t_1 t_2

… Ο κ . Κώστας Κουρταλίδης πάντως προς…

…
Κύριλλος
Κώστας
Λάζαρος
Λάμπρος
Λέων
Λία
…

word, number, symbol
length
first capital, capitalized
end of period
list of names
inside subtitle
-ίδης, -άκης, …
παπα-, χατζη-, …
distance from start of name
κ., κκ.

Vector representation:

< … , … , T, 11, T, F, F, F, T, F, T, F, 1, … , … , … >

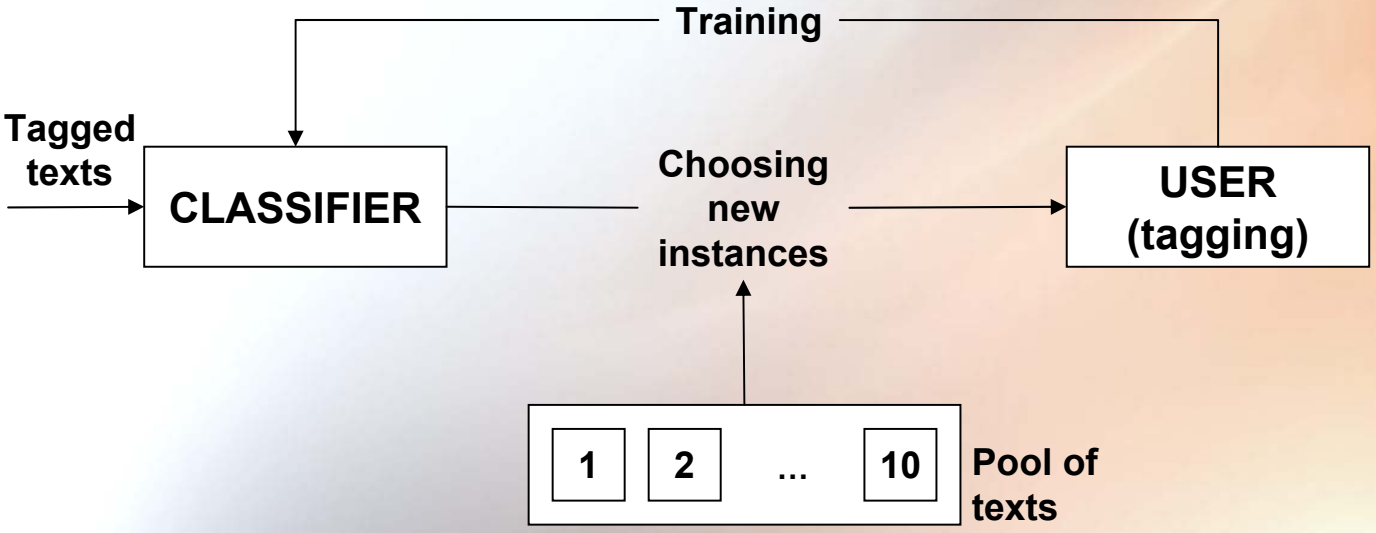
t_{-2} t_{-1} t_0 t_1 t_2

Domain-dependent Features

- **Lists containing words that occur frequently before person tokens**
  - **Different lists for short / long tokens**
  - **Different lists for different distances (immediately before, up to 7 tokens before)**
- **Lists constructed during training**

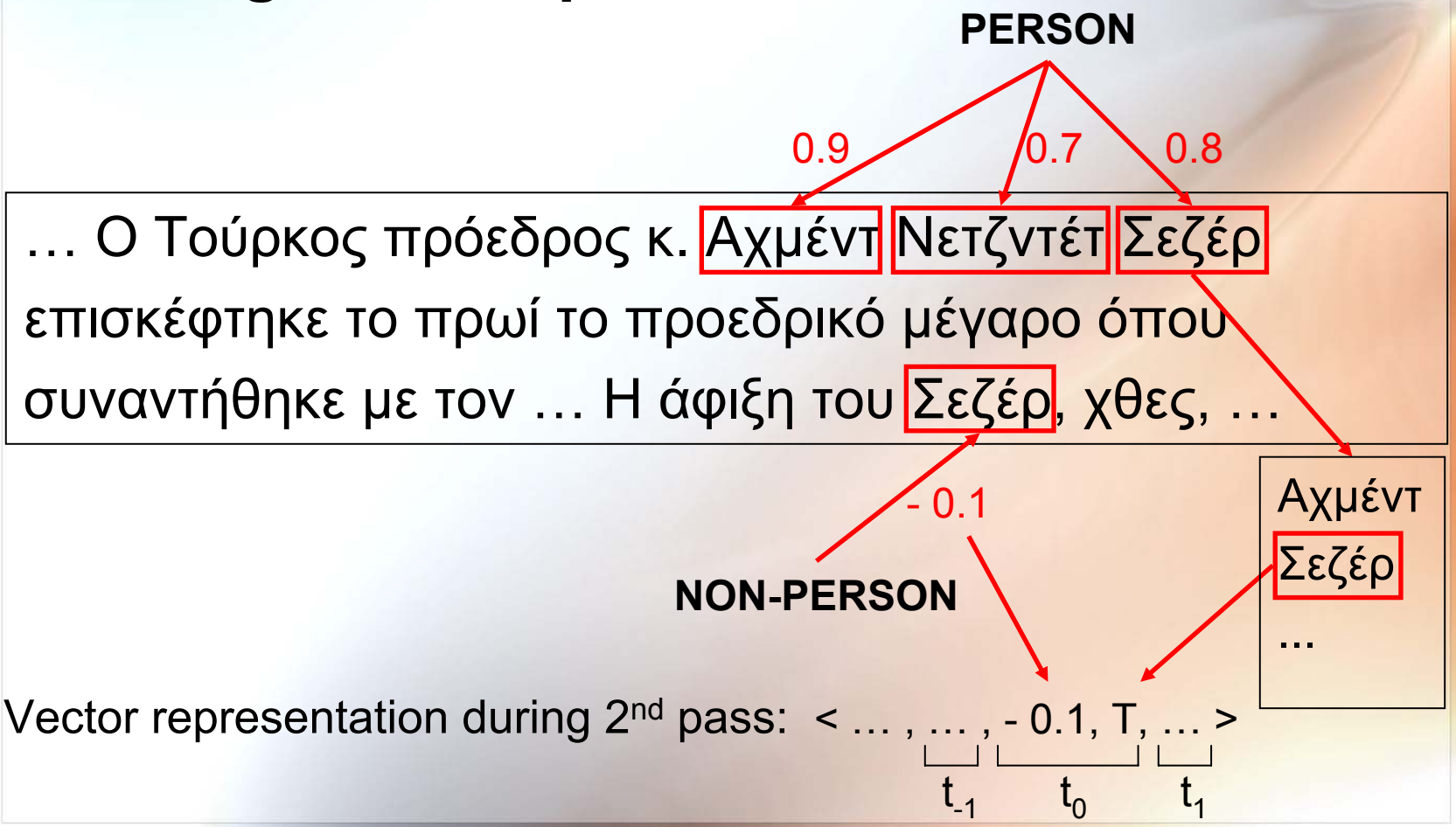
Active Learning

- **Present for human annotation only training instances the system expects will be highly informative.**
  - **Criterion: select instances close to the SVM's hyperplane (maximum uncertainty)**



Second Pass

- **Main idea: Check if any tokens were tagged elsewhere in the same document as person names with high confidence during the first pass.**



Corpora

- 1st collection
  - Newspaper articles (“To Vima”, “Ta Nea”)
  - Topics: politics, finance, sports
  - 400 articles (~331,000 tokens)
    - Person tokens: 4,797
    - Temporal expression tokens: 1,563
  - Tagged using MUC-7 rules
  - Preprocessed to remove HTML tags
    - Header and paragraph tags are kept to determine titles and ends of periods.
- 2nd collection
  - 715 short financial articles (~205,000 tokens)
    - Person tokens: 1,046
    - Temporal expression tokens: 1,244
  - Created during the “MITOS” project

Evaluation

- **Evaluation measures**

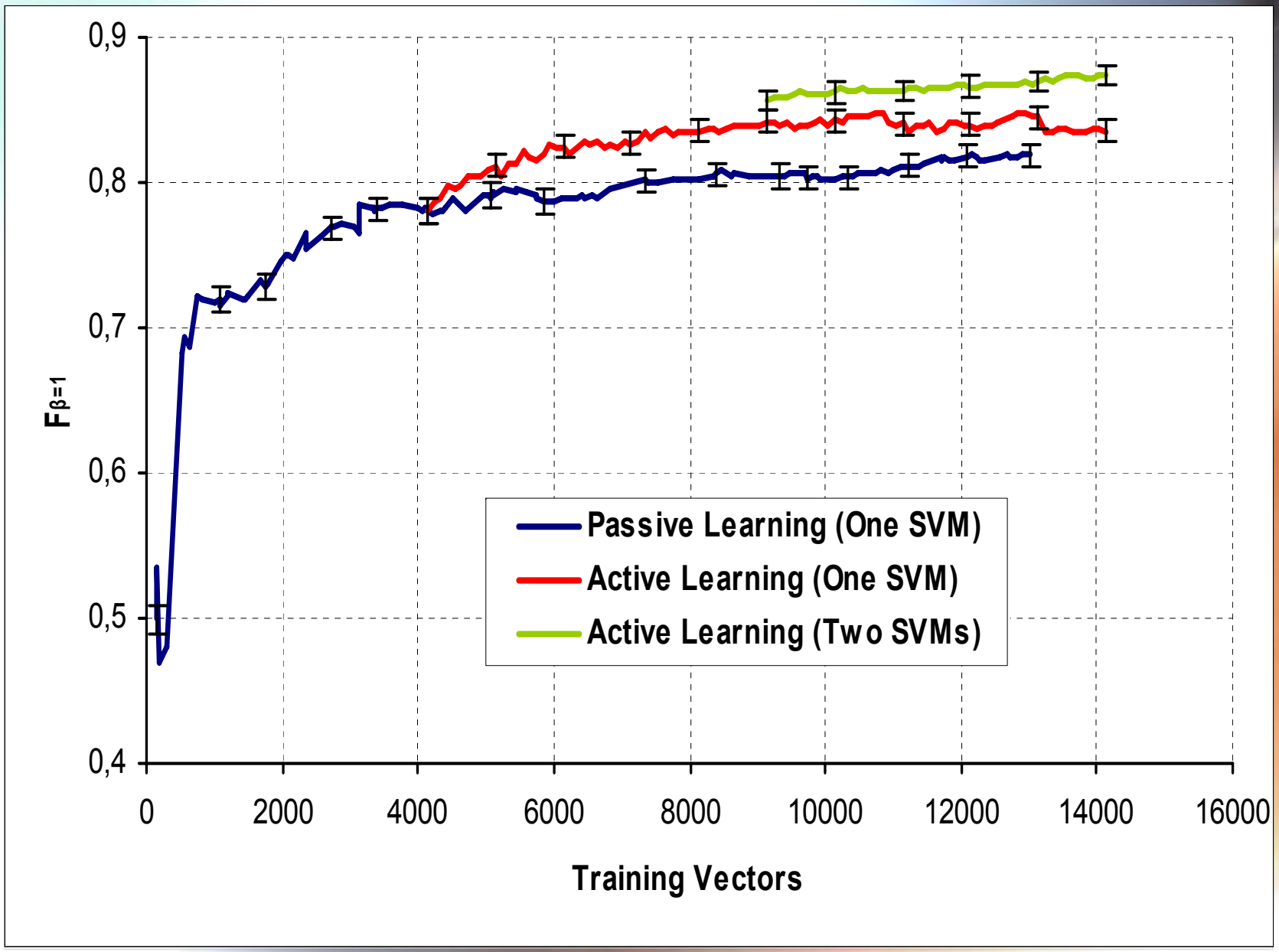
$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

  - TP: #tokens correctly classified as named-entity
  - FP: #tokens wrongly classified as named-entity
  - FN: #tokens wrongly classified as non-named-entity

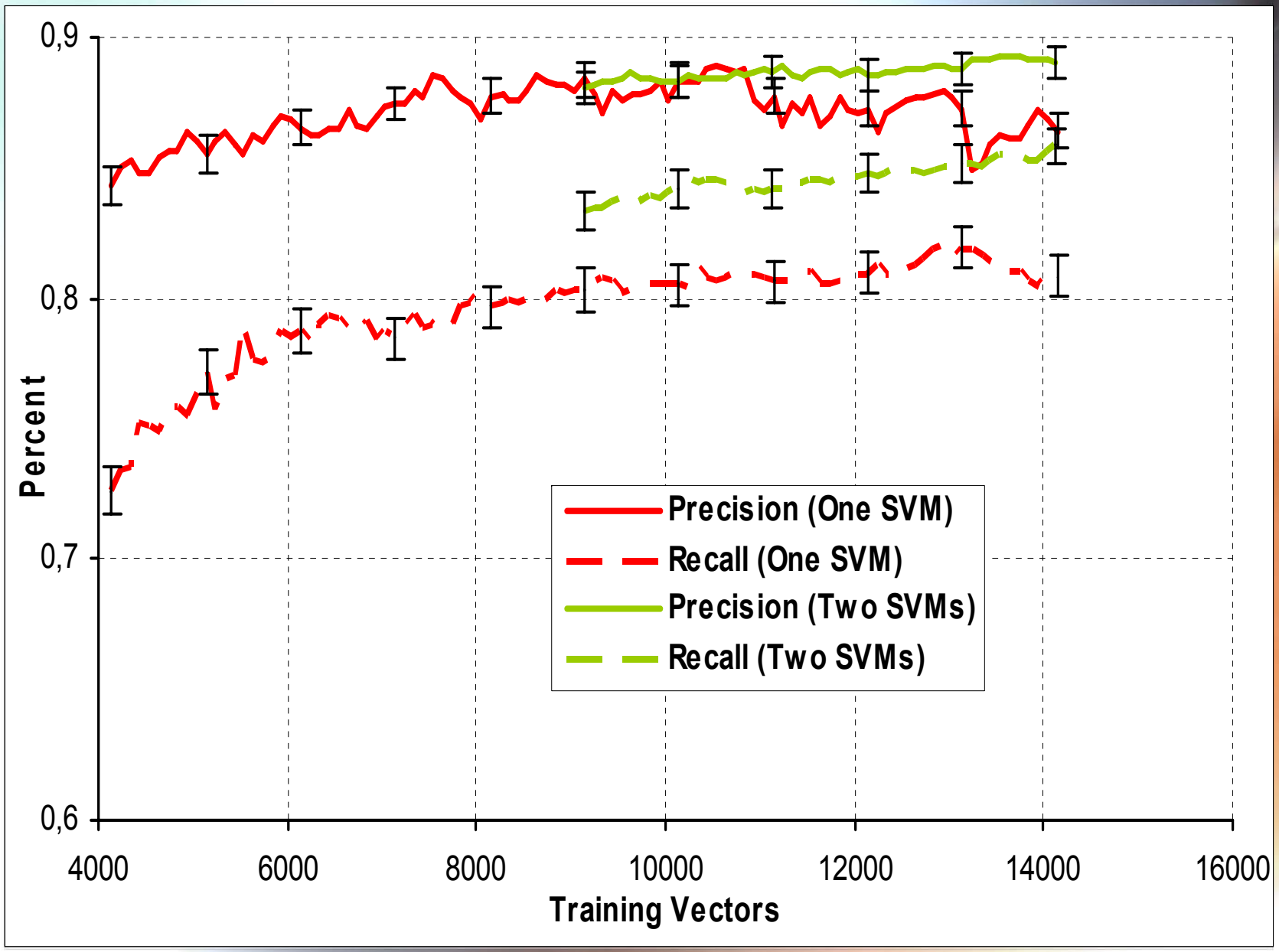
$$\text{F-measure} = \frac{(\beta^2+1) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}, \quad \beta = 1$$

- **Confidence intervals**
  - confidence level 0.99

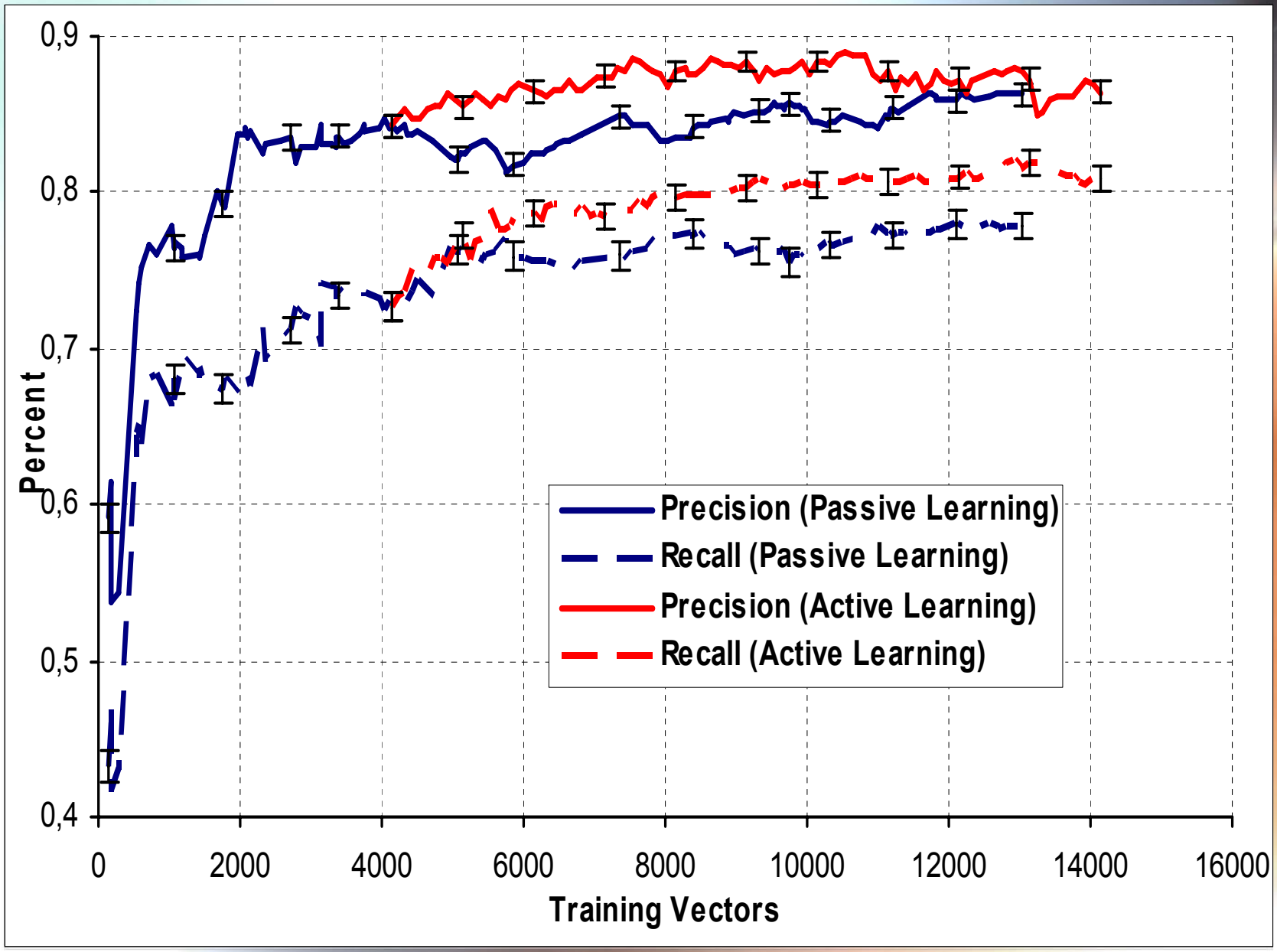
F-measure


0,9
0,8
0,7
0,6
0,5
0,4
Fβ=1
0
2000
4000
6000
8000
10000
12000
14000
16000
Training Vectors
Passive Learning (One SVM)
Active Learning (One SVM)
Active Learning (Two SVMs)

First – Second Pass



Passive – Active Learning



Evaluation results

- **Temporal expression recognition**
  - **10-fold cross-validation**

corpus	precision (%)	recall (%)	$F_{\beta=1}$ (%)
general articles	96.62	92.95	94.75
financial articles	97.59	95.35	96.46

- **Person name recognition**

corpus	methods (training inst.)	precision (%)	recall (%)	$F_{\beta=1}$ (%)
general articles	1 SVM, passive (13K)	86.29 ±0.69	77.91 ±0.83	81.89 ±0.77
	1 SVM, active (9.1K)	88.39 ±0.64	80.33 ±0.79	84.17 ±0.73
	2 SVMs, active (9.1, 14K)	89.06 ±0.62	85.83 ±0.69	87.42 ±0.66
financial articles (cross-validation)	1 SVM, passive (8.8K)	94.96 ±1.28	88.95 ±1.83	91.86 ±1.60
	2 SVMs, passive (8.8, 8.8K)	95.76 ±1.18	91.05 ±1.67	93.34 ±1.46

Conclusions

- **Evaluation of temporal expression recognition**
  - **F-measure**
    - **general articles 94,75%**
    - **financial articles 96,46%**
- **Evaluation of person name recognition**
  - **Sure-fire rules decrease the category imbalance**
  - **F-measure**
    - **general articles 87,42%**
    - **financial articles 93,34%**
  - **Active learning**
    - **F-measure increased up to 4,5-5%**
  - **Second pass**
    - **F-measure increased up to 5%**
- **Future work**
  - **New named-entity categories, e.g. organizations, locations**
  - **Apply system to a different domain**
  - **Improve the active learning criterion**
- **Downloadable from:**

  http://www.aueb.gr/users/ion/publications.html

Previous Greek Systems

- **Person name recognition**
  - **Greek financial news**
  - **Hand-crafted rules**
  - **Boutsis et al. (2000)**
    - **Precision 71%**
    - **Recall 71%**
  - **Farmakiotou et al. (2000)**
    - **Precision 88%**
    - **Recall 77%**
    - **F-measure 82%**
  - **Our results (2005)**
    - **Precision 96%**
    - **Recall 91%**
    - **F-measure 93.5%**
  - **No other comparable results**