Graph-based and Lexical-Syntactic Approaches for the Authorship Attribution Task

Notebook for PAN at CLEF 2012

Esteban Castillo, Darnes Vilariño, David Pinto, Iván Olmos, Jesús A. González and Maya Carrillo

September 12, 2012

Index

Traditional Authorship Attribution

- Authorship attribution assumes unique and identifiable writeprints in text.
- The importance of finding the correct features for characterizing the signature or particular writing style of a given author is fundamental

Lexical-syntactic approach: features

1. Phrase level features
   - **Word prefixes**
     - e.g. *ad* $\longrightarrow$ {**ad***vance*, **ad***junct*, **ad***ulterate*}
   - **Word sufixes**
     - e.g. *est* $\longrightarrow$ {*fin***est**, *tough***est**, *bigg***est**}
   - **Stopwords**
     - e.g. {*and*, *the*, *but*, *did*}
   - **Trigrams of PoS**
     - e.g. she:PRP drove:VBD a:DT silver:NN pt:NN cruiser:NN
       $\{(PRP, VBD, DT), (VBD, DT, NN), (DT, NN, NN), (NN, NN, NN)\}$
2. Character level features
   - **Vowel combination**
     - e.g. *influential* $\longrightarrow$ *iueia* $\longrightarrow$ *iuea*
   - **Vowel permutation**
     - e.g. *influential* $\longrightarrow$ *iueia*

Lexical-syntactic approach: text representation

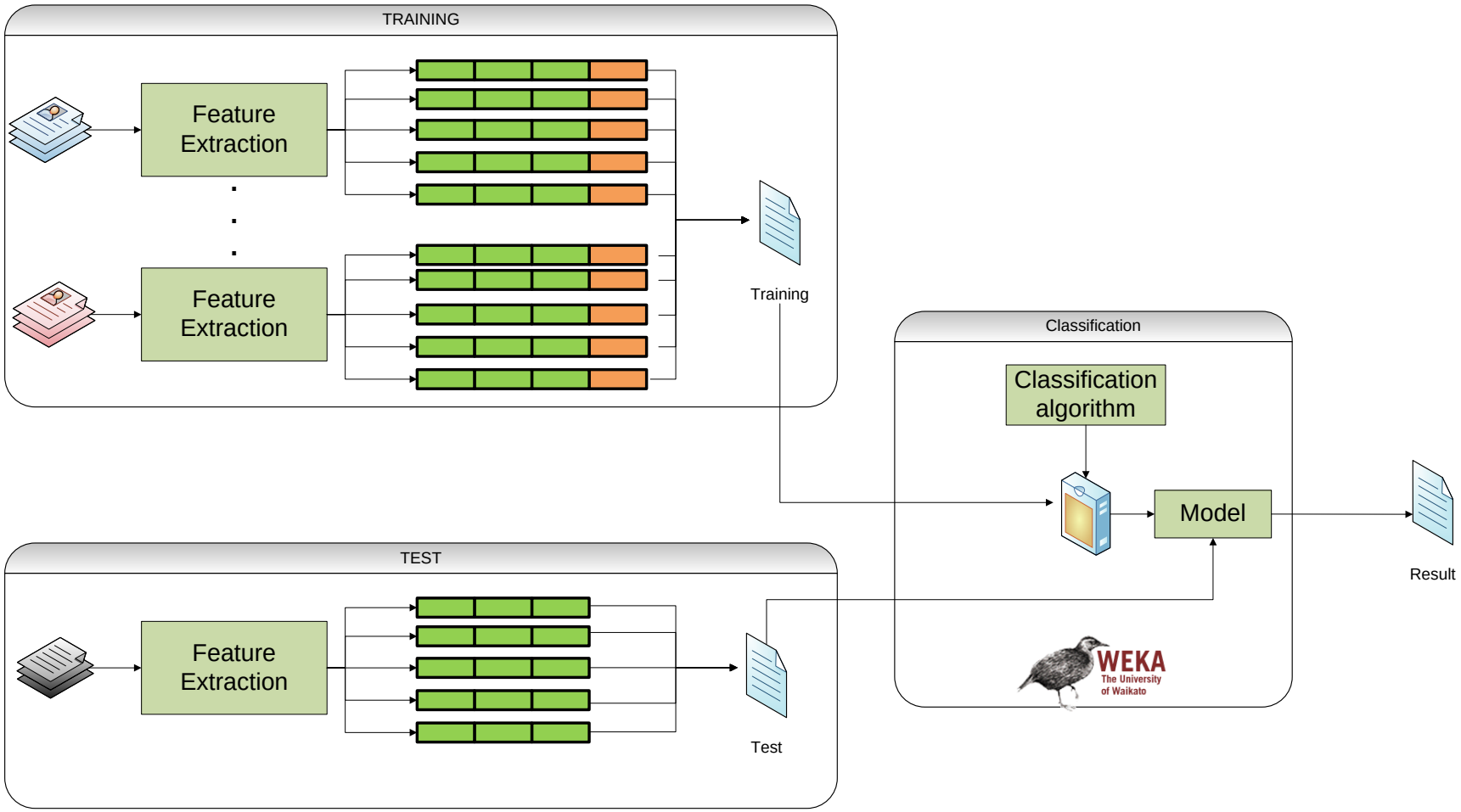
- Training stage:

$$(\underbrace{x_1, x_2, x_3, \ldots, x_s}_{Feature\ 1}, \cdots, \underbrace{y_1, y_2, y_3, \ldots, y_m}_{Feature\ n}, C)$$

- Testing stage:

$$(\underbrace{x_1, x_2, x_3, \ldots, x_s}_{Feature\ 1}, \cdots, \underbrace{y_1, y_2, y_3, \ldots, y_m}_{Feature\ n})$$

Lexical-syntactic approach: Classification process

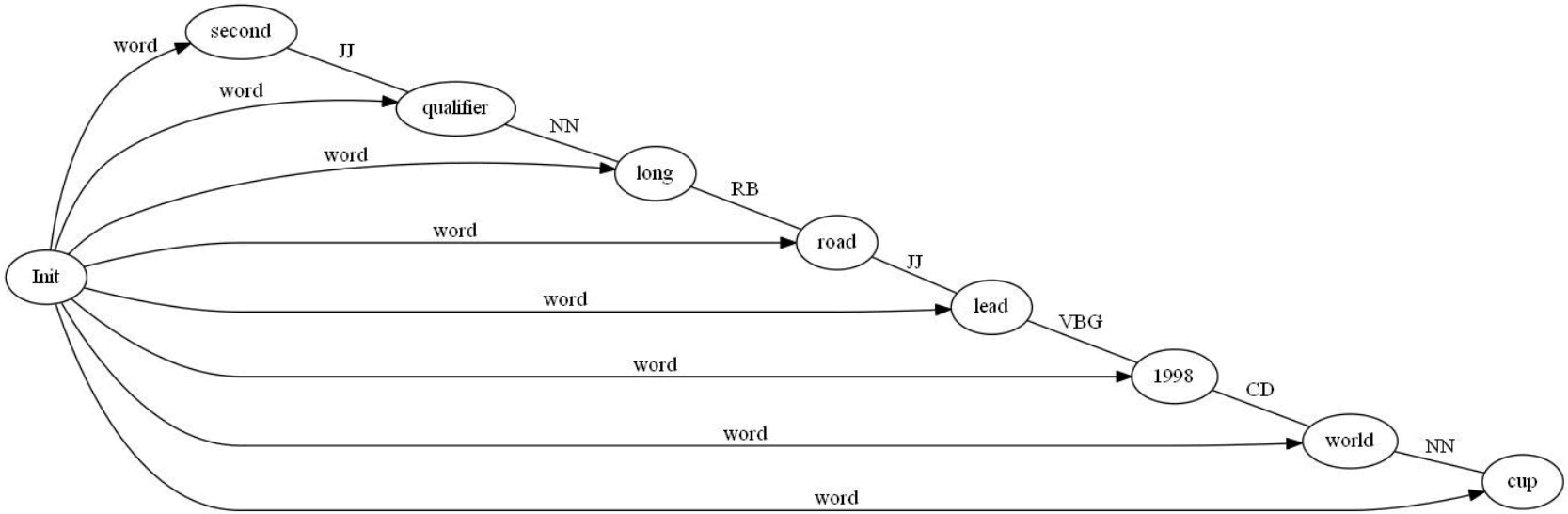


Graph-based approach: features

- In this approach, a graph based representation is considered.
- Each text paragraph is tagged with its corresponding PoS tags with the TreeTagger tool.
- Each word is stemmed using the Porter stemmer.
- In the graph representation each vertex is considered to be a stemmed word and each edge is considered to be its corresponding PoS tag.
- The word sequence of the paragraphs to be represented is kept.
- Once each paragraph is represented by means of a graph, we apply a data mining algorithm called **SUBDUE** in order to find the most representative words of an author

Graph-based approach: example

- "second qualifier long road leading 1998 world cup".



Graph-based approach: text representation

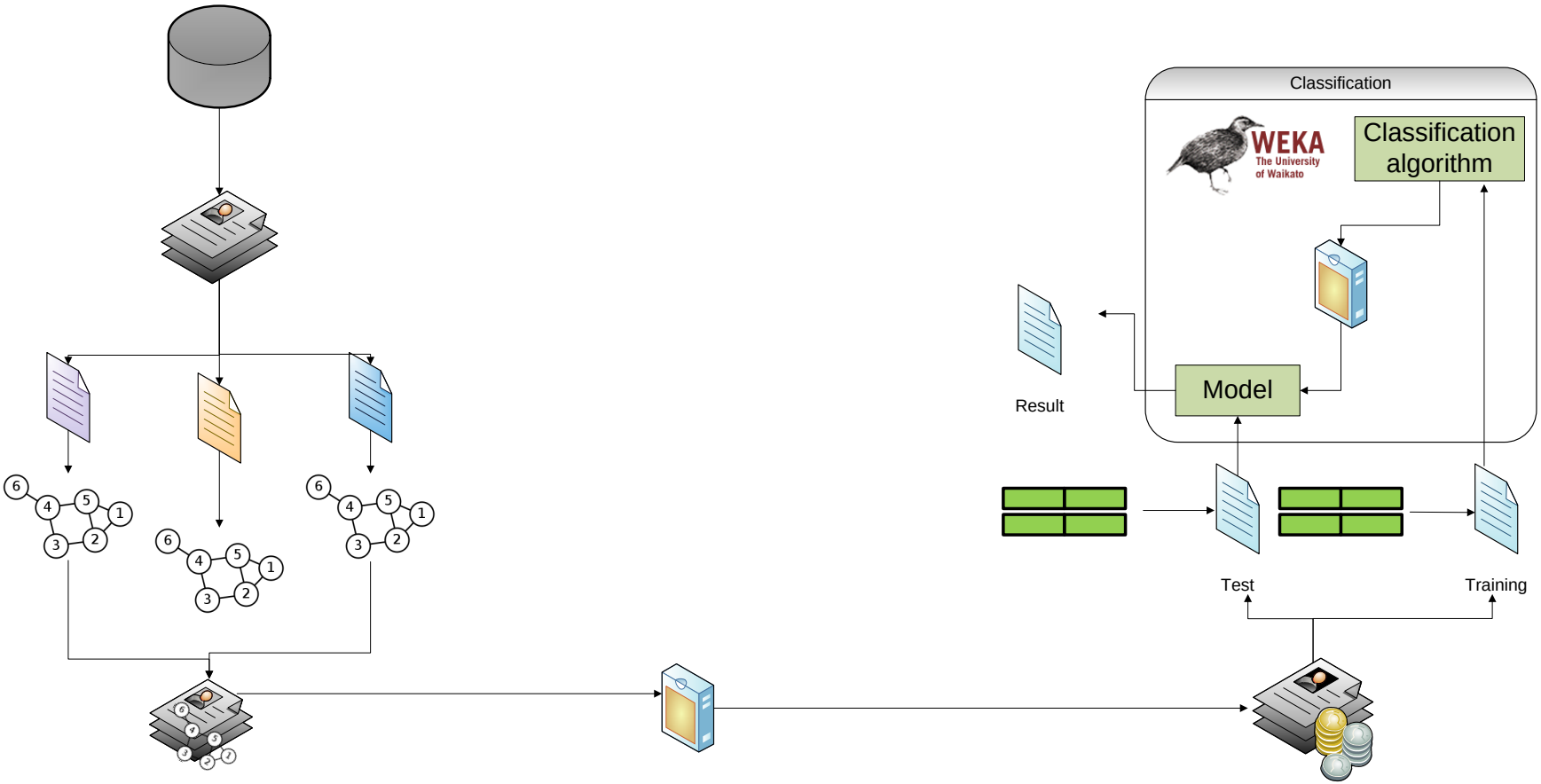
- Training stage:

$$D = (\underbrace{x_1, x_2, x_3, \ldots, x_n}_{Words\ obtained\ from\ SUBDUE}, C)$$

- Testing stage:

$$D = (\underbrace{x_1, x_2, x_3, \ldots, x_n}_{Words\ obtained\ from\ SUBDUE})$$

Graph-based approach: Classification process



Experimental settings

- For SUBDUE we extract the 30 most representative words
- For the problems A, B, C, D, I and J we used WEKA's implementation of SVMs
  - Kernell = polynomial mapping
- For the problems E and F, we used WEKA's implementation K-means clustering method
  - K = 2,3 or 4 authors

Results

Results obtained in the traditional sub-task

Task	A correct/A%	B correct/B%	C correct/C%	D correct/D%	I correct/I%	J correct/J%
Graph-based approach	5/83.333	6/60	5/62.5	4/23.529	8/57.142	13/81.25
Lexical-syntactic approach	4/66.666	3/30	2/25	6/35.294	10/71.428	7/43.75

Results obtained in the clustering sub-task

Task	E correct/E%	F correct/F%
Graph-based approach	68/75.555	43/53.75
Lexical-Syntactic approach	61/67.777	51/63.75

Concluding remarks

1. Lessons learned
   - The lexical-syntactic feature approach helped to represent the writing style
   - the graph-based representation obtained a better performance than the other one. However, more investigation on the graph representation is still required
2. Current work
   - Other data sets and tasks
   - Still more lexical-syntactic features to design and use
   - Understand better the role of the Graph representation
   - Experiment with different graph based text representations that allow us to obtain much more complex patterns.

Thank you for your attention!