A Graph Based Authorship Identification Approach

Helena Gómez Adorno[1], Grigori Sidorov[1], David Pinto Avendaño[2], Ilia Markov[1]

Helena Gómez-Adorno
helena.adorno@gmail.com

[1] Natural Language and Text Processing Lab
Center for Computing Research
Instituto Politécnico Nacional
Mexico City, Mexico

[2] Information Retrieval Lab
Faculty of Computer Science
B. Universidad Autónoma de Puebla
Puebla, Mexico

1. Introduction

- Intrinsic Approach for Authorship Identification.
- Integrated Syntactic Graphs (ISG) for representing texts **[1]** .
- Integration of various levels of the language description.
- Feature extraction based on shortest paths traversal.
- No external documents needed

2. Integrated Syntactic Graph

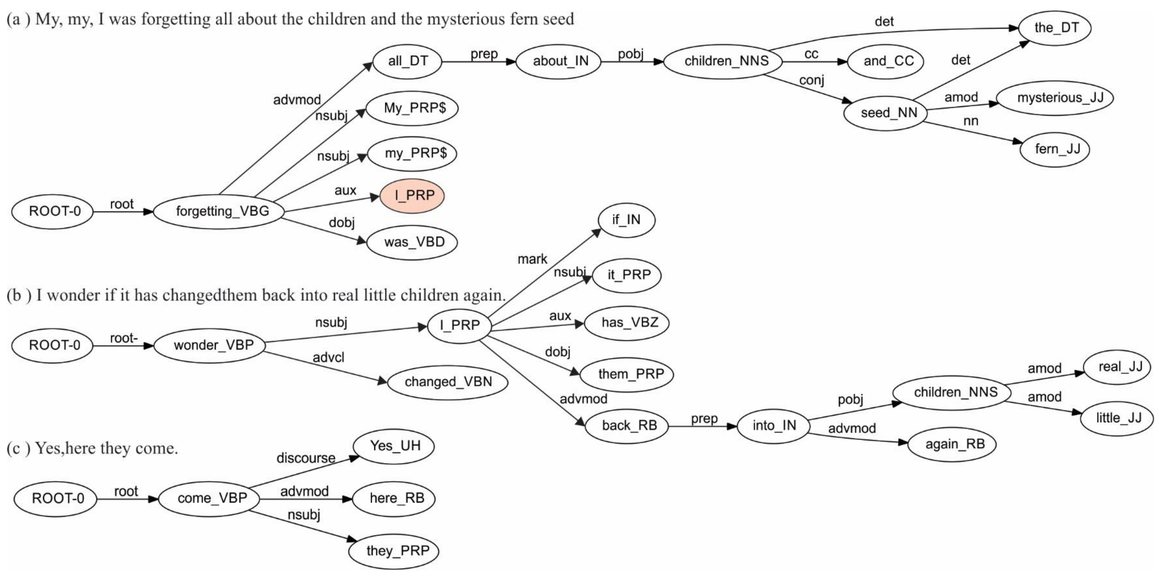


Figure 1: Graph representation of the first three sentences of a text using words, PoS tags and dependency tags.

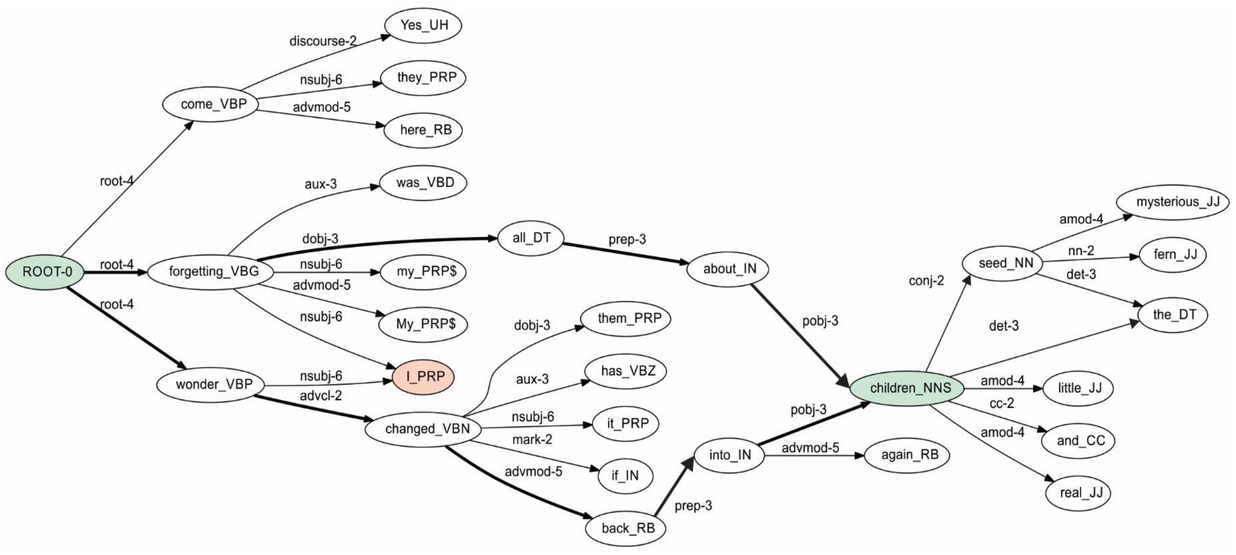


Figure 2: Graph representation of a paragraph using words, PoS tags and dependency tags and nodes frequency information

2.1 Feature Extraction from ISG's

- Considering **Figure 2**, the minimum path from the node ROOT-0 to the node children_NNS will have the following features :
  - **Lexical level:** forgetting, all, about, children.
  - **Morphological level:** VBG, DT, IN, NNS.
  - **Syntactical level:** dobj, prep, pobj.
- For the construction of a vector space model representation of the document, we consider each path as a vector of linguistic elements with numeric values (frequencies).
- For the pair *(ROOT-0, children_NNS)* the shortest path is:
  ROOT-0, forgetting_VBG, all_DT, about_IN, children_NNS.
- A Path-Feature matrix is built with the linguistic information of each path.

Table 1. Vector representation of a document based on shortest paths

Path	Lexical Features				Morphological Features				Syntactic Features			
	forgetting	all	...	I	VBG	DT	...	PRP	dobj	nsubj	...	pobj
ROOT-0 **to** children_NNS	1	1	...	0	1	1	...	0	1	0	...	1
ROOT-0 **to** I_PRP	1	0	...	1	1	0	...	1	0	1	...	0
ROOT-0 **to** I_PRP	0	0	...	1	0	0	...	1	0	1	...	0
ROOT-0 **to** them_PRP	0	0	...	0	0	0	...	1	1	0	...	0

3. Similarity Calculation

- An unknown author's graph D_1 and a known author's graph D_2 are represented by m feature vectors:

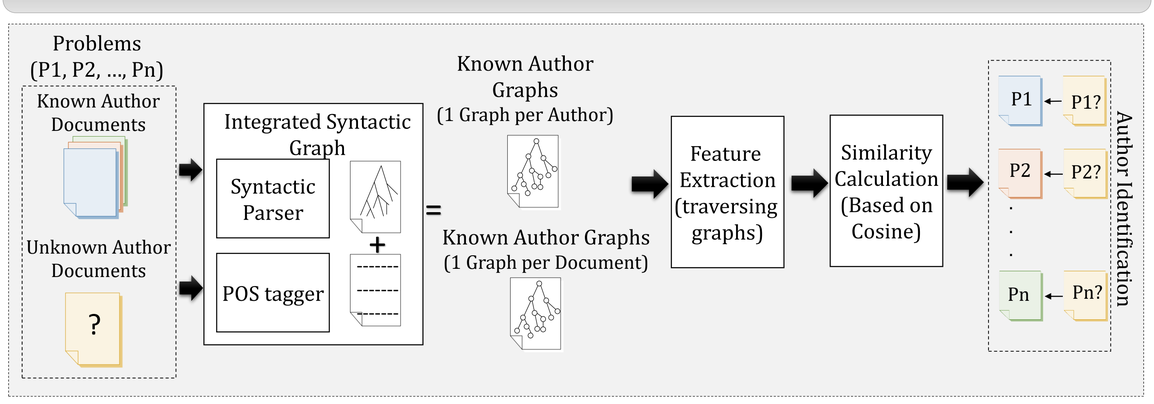
$$D_1 = \{\overrightarrow{f_{D_1,1}}, \overrightarrow{f_{D_1,2}}, \dots, \overrightarrow{f_{D_1,m}}\} \qquad D_2 = \{\overrightarrow{f_{D_2,1}}, \overrightarrow{f_{D_2,2}}, \dots, \overrightarrow{f_{D_2,m}}\}$$

- The similarity is calculated as follows:

$$Similarity\ (D_1, D_2) = \sum_{i=1}^{m} Cosine(\overrightarrow{f_{D_1,i}}, \overrightarrow{f_{D_2,i}})$$

Where:
$\boldsymbol{m}$ is the number of different paths that can be traversed in both graphs
$\boldsymbol{f_{D_1,i}}$ features of the document with unknown author
$\boldsymbol{f_{D_2,i}}$ features of the document with known author

4. Authorship Verification Approach



- The system gives an answer for all the problems.
- It uses the probability scores "0" when the document does not correspond to the author of its problem and "1" if the document belongs to the author of its problem.
- If the similarity is greater than a predefined threshold, then the answer is "1".
- If the similarity is lower than the predefined threshold, then the answer is "0".

5. Results

Table 2. Results obtained for the different languages

Language	AUC	C@1	Final Score	Runtime
English	0.53	0.53	0.28	07:36:58
Spanish	0.53	0.53	0.28	00:50:40
Dutch	0.63	0.63	**0.39**	83:58:15
Greek	0.59	0.59	0.35	00:09:21

Table 3. Ranking for the different languages at Pan 2015

Language	Ranking
English	12/18
Spanish	14/18
Dutch	8/18
Greek	12/18

6. Future Work

- Calculate a confidence score for the answers, instead of answer only "1" and "0" as we did in this version of the system.
- Determine the best configuration of the graph representation to be used for a given corpus.
- Evaluate the performance of the soft cosine measure [2] for this task

References

[1] Pinto, D., Gómez-Adorno, H., Ayala, D.V., Singh, V.K.: A graph-based multi-level linguistic representation for document understanding. *Pattern Recognition Letters* **41**, 2014, 93-102.

[2] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas* 1**8**(3), 2014.