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

(a) My, my, I was forgetting all about the children and the mysterious four pennies.

(b) I wonder if it has changed for the kids in this little city again.

(c) You're they some.

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 the ISG

- 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

<table>
<thead>
<tr>
<th>Path</th>
<th>Lexical Features</th>
<th>Morphological Features</th>
<th>Syntactic Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>forgetting, all</td>
<td>VBG, DT, IN, NNS</td>
<td>dobj, prep, pobj</td>
</tr>
<tr>
<td>ROOT-0 to children, NNS</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ROOT-0 to L_PPP</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>ROOT-0 to L_PRED</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>ROOT-0 to them, PRP</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

3. Similarity Calculation

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

$$D_1 = \left[ f_{D_1,1}, f_{D_1,2}, ..., f_{D_1,m} \right], \quad D_2 = \left[ f_{D_2,1}, f_{D_2,2}, ..., f_{D_2,m} \right]$$

- The similarity is calculated as follows:

$$\text{Similarity}(D_1, D_2) = \frac{\sum_{i=1}^{m} \cos(f_{D_1,i}, f_{D_2,i})}{m}$$

where:

$m$ is the number of different paths that can be traversed in both graphs

$f_{D_1,i}$ features of the document with unknown author

$f_{D_2,i}$ features of the document with known author

4. Approach to Authorship Identification

- 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

<table>
<thead>
<tr>
<th>Language</th>
<th>AUC</th>
<th>@1</th>
<th>Final Score</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.53</td>
<td>0.53</td>
<td>0.28</td>
<td>07:36:58</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.53</td>
<td>0.53</td>
<td>0.28</td>
<td>00:50:40</td>
</tr>
<tr>
<td>Dutch</td>
<td>0.63</td>
<td>0.63</td>
<td>0.39</td>
<td>83:58:15</td>
</tr>
<tr>
<td>Greek</td>
<td>0.59</td>
<td>0.59</td>
<td>0.35</td>
<td>00:09:21</td>
</tr>
</tbody>
</table>

Table 3. Ranking for the different languages at Pan 2015

<table>
<thead>
<tr>
<th>Language</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>12/18</td>
</tr>
<tr>
<td>Spanish</td>
<td>14/18</td>
</tr>
<tr>
<td>Dutch</td>
<td>8/18</td>
</tr>
<tr>
<td>Greek</td>
<td>12/18</td>
</tr>
</tbody>
</table>

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
