Using Explicit Semantic Analysis for Linking in Multi-lingual Document Collections

ESADocLink is a system that automatically suggests links for documents in a multi-lingual document collection.

Motivation:
- People prefer structured collections, like Internet or Wikipedia.
- The links help me not to oversee a document when researching a topic!
- I can browse the collection by navigating through links!

What other documents in there are on the topic of Apple?
Is there a document which discusses this issue in more depth?
A piece of fruit mentioned in an article on computer history. I don't understand. What is that?

Once upon a time, there was a flat document collection...
... but people don't like such collections, because they are difficult to navigate.

Flat document collection
like the majority of document collections today

ESADocLink is a system that automatically suggests links for documents in a multi-lingual document collection.

1. Document Discovery
   Select documents from the document collection that should be linked with the input document (different approaches):
   a] semantically similar documents
   b] from existing links in the most semantically similar documents
   c] by average semantic representation
   d] dimensions of the ESA vector

2. Link Placement
   - look for a position in the input document, which should link to the target document
   - searches the document for the possible matches with the title of the discovered documents

3. Link Classification
   - using SVM with the following features:
     - relative link position,
     - link frequency,
     - generality,
     - ESA similarity (input vs. target, link text vs. target document, link text vs. target document title)

Concepts

<table>
<thead>
<tr>
<th>Concept</th>
<th>ESA vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tablet</td>
<td>0.2</td>
</tr>
<tr>
<td>Screen</td>
<td>0.15</td>
</tr>
<tr>
<td>Forest</td>
<td>0.001</td>
</tr>
<tr>
<td>Toy</td>
<td>0.1</td>
</tr>
</tbody>
</table>

ESA is the core of the system