ENSM-SE at INEX 2012: Basic Experiments

Philippe Beaune, Michel Beigbeder, and Mihaela Juganaru-Mathieu

École Nationale Supérieure des Mines de Saint-Étienne
Institut Henri Fayol
158 cours Fauriel, F 42023 SAINT ETIENNE CEDEX 2, France
beaune@emse.fr, mbeig@emse.fr, mathieu@emse.fr

1 Introduction

Our objective in the INEX 2012 campaign was to integrate the semantic tags and the linked data in our proximity retrieval model. This model was sucessfully used in previous INEX campaigns and obtained good results, particularly in 2007 with the second place in the Ad Hoc Track Focused Task [1], and in 2010 with the first place in the Ad Hoc Track Relevant in Context Task [2]

Though we had several discomfitures with the collection because i) there were several versions of the collection, the last one available at the end of June, one week before the initial run submission deadline, ii) the different versions were difficult to follow because they were not clearly identified, iii) not every documents were well formed according to the XML format, iv) the provided DTD gives little information on the actual structure and its semantics, v) the documents contains many semantic annotations but the underlying ideas used to generate them are not documented making them difficult to apprehend. We present in section 2 how we processed the documents to alleviate the problems with the DTD.

Thus we only have been able to do some basic experiments presented in section 3. In section 4 we present our work in progress.

2 Collection preparation

The collection comes with 3 164 040 documents, of which 4 749 are not well formed according to the XML format. We deleted these documents in our experiments as they only represent 0,15% of the collection.

Structure was extremely difficult to apprehend with the provided DTD (wikipedia-lod-xml.dtd) because almost every elements can contain any other one. Here is a small extract of this DTD:

```
<!ELEMENT wikipedia ( heading | list | paragraph | table | hr | list |
preblock )* >

<!ELEMENT heading ANY >
<!ATTLIST heading level CDATA #IMPLIED >

<!ELEMENT list ( listitem+ ) >
```

```
<!ATTLIST list type NMTOKEN #REQUIRED >

<!ELEMENT listitem ANY >

<!ELEMENT paragraph ANY >
```

Some XML elements (such as `wikipedia` and `list`) are well defined because they could only contain a small number of meaningful elements. But 46 of the 70 XML tags defined in this DTD can contain any content, such as `heading`, `listitem` and `paragraph`.

With this DTD the following extract can be a part of a valid document:

```
[...]
<heading>
   <listitem>
      <paragraph>
          <heading>
[...]
```

where the structure has no sense using the usual meaning of the words *heading*, *paragraph* and so on.

So we decided to build a new collection where each document validates the very simple following DTD:

```
<!ELEMENT article ( title, CDATA ) >
<!ELEMENT title CDATA>
```

Some elements were deleted, for example `yagoproperties` and `dbpediaproperties`. For the other elements we only kept their textual content. We also ignored all the attributes except the attribute `@name`, whose value was kept as text. This operation was done with `xsltproc` and processing the whole collection lasted more than 17 hours.

We also tried to use `TreeTagger`[3] but it was too slow to process the whole collection because each document needed around one second to be processed.

Finally, the collection and its very simple structure was indexed with `zettair`[1] with the light stemmer on, lasting 40 minutes.

3 Runs

Three runs were allowed for participants in INEX 2012. Two of our runs were produced with *zettair*, the first one, *Emse-085*, used a language model with a Dirichlet smoothing. The second one, *Emse-086*, used the well known BM25 model with $k_1 = 1.2$, $k_3 = +\infty$ and $b = 0.75$. Both these runs were produced within 30 seconds for the 140 queries.

The third run, *Emse-087* used our proximity model developed for the previous INEX campaigns [4, 5], and its execution needed 2 minutes and 45 seconds.

For the present we do not have the assessments so no evaluation was performed.

[1] http://www.seg.rmit.edu.au/zettair/

4 Perspectives

4.1 Proximity model

Our proximity model works with the following type of structured documents:
$document \leftarrow (part)^+$
$part \leftarrow text$
$part \leftarrow (part)^+$
$part \leftarrow title \oplus (part)^+$
For plain text our model computes a score based on a fuzzy neighbouring parameterized function. For a document composed of a concatenation of parts, the score is the sum of the part scores. For a document/part with a title, title words are considered as close to any word of the part content.

4.2 First choice

The provided DTD doesn't permit us to easily construct a collection fulfilling the above description.The title of the documents was easy to extract, but as the part titles and the parts themselves are not nested, extracting these titles to insert them in their corresponding part is not possible in XSLT [6]. So we considered the XML documents as:
$document \leftarrow title \oplus text$
and we applied our model in this simplified case.

4.3 Future works

We detected that the tag `heading` could be the title of parts, but the parts themselves are not explicit and clearly delimited. We will construct a new collection fulfilling our document model using a high level programming language using the library `libxml` and build the nesting based on the assumption that the attribute `@level` of the tag `heading` indicates the actual nesting.

We will also consider the tags `yagoproperties` and `dbpediaproperties` as parts of the newer documents. This work is in progress.

References

1. Fuhr, N., Kamps, J., Lalmas, M., Malik, S., Trotman, A.: Overview of the inex 2007 ad hoc track. In Fuhr, N., Kamps, J., Lalmas, M., Trotman, A., eds.: INEX. Volume 4862 of Lecture Notes in Computer Science., Springer (2007) 1–23
2. Arvola, P., Geva, S., Kamps, J., Schenkel, R., Trotman, A., Vainio, J.: Overview of the inex 2010 ad hoc track. In Geva, S., Kamps, J., Schenkel, R., Trotman, A., eds.: INEX. Volume 6932 of Lecture Notes in Computer Science., Springer (2010) 1–32
3. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing. (September 1994)
4. Beigbeder, M., Imafouo, A., Mercier, A.: ENSM-SE at INEX 2009 : Scoring with proximity and semantic tag information. **6203** (2009) 49–58

5. Beigbeder, M.: Focused retrieval with proximity scoring. In Shin, S.Y., Ossowski, S., Schumacher, M., Palakal, M.J., Hung, C.C., eds.: SAC, ACM (2010) 1755–1759
6. Møller, A., Olesen, M.O., Schwartzbach, M.I.: Static validation of xsl transformations. ACM Trans. Program. Lang. Syst. **29**(4) (August 2007)