CLEF–IP 2011
Track Guidelines

1 Introduction

The CLEF–IP track was launched in 2009 to investigate IR techniques for patent retrieval and it was part of the CLEF 2009 evaluation campaign. In 2011, the track continues as a benchmarking activity of the CLEF 2011 conference.

The track utilizes a collection of more than 1.3 million patent documents derived from EPO sources. The collection covers English, French and German with at least 150,000 documents in each language.

There are five tasks in the 2011's track.

- Prior Art Candidate Search: Find patent documents that are likely to constitute prior art to a given patent application.
- Classification: Classify a given patent document according to the IPC system, up to the subclass level.
- New: Refined Classification: Classify a given patent document up to the group/subgroup level, when the subclass is given.
- New: Image-based Document Retrieval: Find patent documents or images relevant to a given patent document containing images.
- New: Image-based Classification: Categorize given patent images into pre-defined categories of images (such as graph, flowchart, drawing, etc.).

This document describes the first three tasks. For more information on the new image related tasks refer to the homepage http://www.ir-facility.org/clef-ip.

2 Target data

The target data set contains all EPO documents that have an application date previous to 2002. (more than 2.5 Million patent documents constituting more than 1Million patents). In addition for EuroPCT Applications we also added the corresponing patent documents published by the WIPO (more than 400,000 documents). The same as in the last years, the Test Collection Corpus is delivered to the participants "as is", without merging the documents related to the same patent into one document. Each patent is identified by a unique patent number. Corresponding to each patent is a directory containing the patent documents related to that patent. The layout is CC/nnnnnn/nn/nn/nn/*.xml.

For example, to patent EP 0981201 corresponds the directory containing files EP-0981201-A2.xml, EP-0981201-A3.xml, and EP-0981201-B1.xml:

```
> pwd
EP/000000/98/12/01
> ls
EP-0981201-A2.xml EP-0981201-A3.xml EP-0981201-B1.xml
```

Patent documents

In general, one patent (identified by a unique patent number) corresponds to several patent documents generated at different stages of the patent's life-cycle. Each stage is denoted by a kind code together with a version number[1]. The most common kind codes[2] are the following:

A1	publication of application with search report
A2	publication of application without search report
A3	publication of search report
A4	supplementary search report
A8	corrected title page of an EP–A document
A9	complete reprint of an EP–A document
B1	granted patent
B2	granted patent after modification
B8	corrected front page of an EP–B document
B9	complete reprint of an EP–B document

The "B" kind codes mark granted patents, and they necessarily have been earlier published as an "A1" or "A2" document.

3 The Prior Art Candidate Search Task

The Prior Art Candidate Search Task (PAC) constist in finding parent documents in the target collection that may invalidate a given patent application. In intellectual property language: find documents that may constitute prior art for a topic patent. The topic document is a patent application document, A1 or A2, where the citation information was removed. This years topics are more equally distributed over the different languages, with more than 1000 topics for english, french and german each. A PAC topic file contains a concatenation of topics. The structure of one topic is as follows:

[1] For the EP patents, documents at different stages have the same numeric identifier. For other patent offices this is not always the case. For example, the patent document US-6689545-B2 represents a US granted patent with its application document publication number US-2003011722-A1

[2] A full list of kind codes can be found at http://tinyurl.com/EPO-kindcodes

```
<topic>
    <num>EP-1221372-A2</num>
    <narr>Find all patents in the collection that potentially invalidate
          patent application EP-1221372-A2.</narr>
    <file>EP-1221372-A2.xml</file>
</topic>
```

where `<num>` contains the unique topic identifier consisting of the patent number, which itself contains a country code (always `EP` in this data set), a seven-digit number and the kind code (`A1`, `A2`). The `<file>` tag contains he name of the XML file.

4 The Classification Tasks

This year there are two classification tasks on a set of 3,000 topic documents comprising 1,000 English-, German- and French-language documents each. The topic files are named `CLS<n>_<lang>_topics.txt` with `n` being the task number, and `<lang>` being `en`, `de` or `fr`. Note that the classification system used by the whole clef-ip collection is IPC-R and later. The results too should be IPC-R codes.

4.1 Task 1

The goal for Task 1 (CLS1) is to classify a given topic document according to the International Patent Classification (IPC)[3] on SUBCLASS level.

The topic structure is as follows (example):

```
<topic>
  <num>CLS1_EP-1469052-A1</num>
  <narr>Classify patent document EP-1469052-A1 according to the
    IPC system.</narr>
  <file>EP-1469052-A1.xml</file>
</topic>
```

where `<num>` contains the unique topic identifier consisting of the prefix `CLS1_` and the patent number, which itself contains a country code (always `EP` in this data set), a seven-digit number and the kind code (`A1`, `A2`). The `<file>` tag contains he name of the XML file.

4.2 Task 2 - Refined Classification

The goal for Task 2 (CLS2) is to classify a given topic document with given SUBCLASS on SUBGROUP level.

The topic structure is similar to Task 1 (example):

```
<topic>
  <num>CLS2_EP-1674081-A1_A61K</num>
  <narr>Reclassify patent document EP-1674081-A1 classified in
    subclass A61K into subgroup.</narr>
```

[3] IPC documentation entry point: http://www.wipo.int/classifications/ipc/en/

```
  <subclass>A61K</subclass>
  <file>EP-1674081-A1.xml</file>
</topic>
```

Here, the topic identifier additionally contains the given SUBCLASS, since one document can have multiple classifications, i.e. it can fall into multiple SUBCLASSES. The total number of topics for Task 2 is 4,934. The SUBCLASS is also given in the tag `<subclass>`.

5 Training Set

Similar to last year, we have released a set of topics, together with their relevances for training purposes. The participants can use this set to train and tune their systems. The CLEF–IP 2011 training set contains documents and relevance assessments for 300 topics similar to the PAC search task. The relevance assessments are automatically extracted from patent citations information, and have the following format:

```
EP-0000001-A1    EP-7654321   1
EP-0000001-A1    EP-7654322   2
EP-0000001-A1    EP-7654323   1
```

where the first column contains the topic number, the second column contains the relevant patent as identified by the coutry and patent-number, and the last column is the relevance degree (2 being more relevant than 1). We did not compile a training set for the classification task as Participants can use the whole target data of the PAC task to train their classifiers.