Automating Eurovoc Descriptor Assignment

Automating the Assignment of Eurovoc Descriptors to text

Eurovoc Conference 2003
European Parliament
Brussels, Friday, 7 March 2003

Ralf Steinberger & Bruno Pouliquen
European Commission – Joint Research Centre (JRC)
Institute for the Protection and Security of the Citizen (IPSC)

http://www.jrc.it/langtech

Agenda

- Who we are and what we do
- Overview of our approach to Eurovoc indexing
- More detailed explanation of our approach
- Evaluation procedure and results for the automatic assignment
- Some problems we encountered, and possible solutions
- Why I am here: our offer and request

- Automatic Eurovoc descriptor assignment will not replace the work of professional indexers, but could help them.
JRC Sites in Europe

Photograph of the Ispra Site

Joint Research Centre (JRC / CCR / GFS)

- Directorate General (DG) of the European Commission
- >1500 scientists and technicians, ca. 2500 people
- Ispra: ca. 1800 people
- Scientific research and scientific services for DGs
  - Wide range of subjects:
    - nuclear safety
    - environment (alternatives to animal testing, recognition of adulterated wine, (non-)biological food, ...)
    - Security of food and chemical products
    - dependability of information systems
    - ...
Automating Eurovoc Descriptor Assignment

Goal of JRC’s Language Technology work

- **Retrieval** of potentially relevant texts (e.g. from the internet) in a variety of languages, using agent technology: [OSILIA project (2000), IDoRA for OLAF (2002/03), Breaking News – Detection and Visualisation (2003)]

- **Text analysis** and extraction of a variety of information aspects from texts; when possible: language-independent representation of the contents
  - key words (monolingual free indexing terms and cross-lingual Eurovoc descriptors)
  - language of texts
  - references to geographical places (and to dates)
  - (references to people, to products, etc.)
  - summary
  - Calculation of the similarity of texts; find related documents, even across languages
  - clustering and classification of documents

- **Visualisation of the contents**
  - of individual documents in document profiles
  - of whole text collections in document maps
  - of extracted geographical information in maps

Sample Text: Plutonium Smuggling

In the summer of 1994 a suitcase containing plutonium illegally imported into Germany was seized in sensational circumstances at Munich airport in the Federal Republic of Germany. Is the Commission aware of this matter and, if so, when were the Commission and its services, and other European agencies, informed of it? Can the Commission say whether the Joint Research Centre in Karlsruhe was involved, what services it provided for the German police, when it provided them, when the plutonium was seized, and when it was handed over to the Joint Research Centre?

2: Answer given by Mr. Papoutsis on behalf of the Commission (10 January 1996)

The Commission would refer the Honourable Member to its earlier replies to questions about this incident (Written questions 1465/95[1] OJ C 213, 17.8.1995 and 1503/95[2] OJ C 230, 4.9.1995) by Mr. Breyen. The Commission (European safeguards directorate) was alerted by the German authorities in the early afternoon of 10 August 1994 that some material might be seized. In accordance with formal agreements between the Commission and the German government this information was immediately passed by phone to the European Institute for Transuranium Elements (TUJ) at Karlsruhe to ensure that preparations were made to receive any material seized. The seizure was made by the German police, and the TUJ was not involved. Its activities that night were limited to receiving the closed suitcase at its premises in Karlsruhe. Subsequently, the TUJ performed a precise analysis of the material found inside the suitcase to support the investigations carried out by the Member State authorities and to determine as far as possible the source and history of the nuclear material.
Structured Multilingual Display of Monolingual Information

Title: Seizure of plutonium at Munich airport
Author: Ralf Steinberger

Keywords (Occurrence Frequency)
- plutonium, input, multilingual, Federal Republic of Germany

Eurovoc Thesaurus Descriptors
- Seizure of plutonium at Munich airport
- Seizure of plutonium at Munich airport in the summer of 1994
- Federal Republic of Germany

Agenda

- Who we are and what we do
- Overview of our approach to Eurovoc indexing
- More detailed explanation of our approach
- Evaluation procedure and results for the automatic assignment
- Some problems we encountered, and possible solutions
- Why I am here: our offer and request
JRC Approach – Overview

- Rule-based (linguistic) approach would be:
  - \((\text{nuclear OR radioactive}) \text{ AND } (\text{accident OR leak})\) \(\rightarrow\) NUCLEAR ACCIDENT
  - Time-consuming task
  - Rules have to be written separately for each language

- JRC's statistical, associative approach (bag-of-words approach)
  - Identify many (statistically or semantically) related words (associates)
    (Training phase)
  - Assign descriptor if many of its associates are present in text. (Assignment phase)

The JRC Approach in a Nutshell (1)

**EFTA COUNTRIES**

+++SIMPLIFICATION OF FORMALITIES+++
The JRC Approach in a Nutshell (2)

EFTA COUNTRIES

SIMPLIFICATION OF FORMALITIES

council decision of 22 November 1993 concerning the conclusion of the Agreement in the form of an exchange of letter between the European Community and the Republic of Austria, the Republic of Finland, the Republic of Iceland, the Kingdom of Norway, the Kingdom of Sweden and the Swiss Confederation relating to the amendment of the Convention of May on the simplification of formalities in trade in goods.

THE council of the European Union, having regard to the treaty establishing the European Community, and in particular Article 113 thereof, have regard to the proposal from the Commission, whereas Article 11 (2) of the Convention between the European Economic Community and the Republic of Austria, the Republic of Finland, the Republic of Iceland, the Kingdom of Norway, the Kingdom of Sweden and the Swiss Confederation on the simplification of formalities in trade in goods (1) empower the joint committee set up by that Convention to make recommendations for amendment to the Convention.

Whereas the Convention have been amended to allow for the accession of new Party;

Whereas the agreement in the form of an exchange of letter relating to the recommendation should be approved;

Whereas the amendment in question be set out in recommendation No 1/93 of the joint committee;

Have decided as follows:

Article 1. The Agreement in the form of an exchange of letter between the European Community and the Republic of Austria, the Republic of Finland, the Republic of Iceland, the Kingdom of Norway, the Kingdom of Sweden and the Swiss Confederation relating to the amendment of the Convention of 20 May 1987 on the simplification of formalities in trade in goods be hereby approved on behalf of the Community.

The text of the Agreement be attached to this Decision.

Article 2. The President of the Council be hereby authorized to designate the person empowered to sign the Agreement in order to bind the Community.

Done at Brussels, 22 November 1993.

Library Usage vs. JRC Usage

- Automatic assignment is an approximation, a "best guess".
- Achievable quality is clearly lower than that of human assignment. Libraries need high quality for indexing and retrieval.
- Requirements differ.

- JRC also wants to
  - index, retrieve, and give cross-lingual information access.
  - We index documents that would otherwise not be indexed at all.
  - Cross-lingual document similarity calculation
  - Multilingual classification
  - Multilingual clustering of documents
  - Multilingual document maps
  - Subject-specific summarisation
Identify Statistically Salient Words in Text

- Compare the word frequency (lemma) in a document (TF) with an ‘expected’ / average word frequency (reference corpus frequency RCF)
- using the statistical log-likelihood test (Dunning 1993), or others.
- **Text length**: 300 words; **Reference corpus length**: 100 million words

<table>
<thead>
<tr>
<th>Lemma</th>
<th>TF</th>
<th>RCF</th>
<th>Keyness</th>
</tr>
</thead>
<tbody>
<tr>
<td>tui</td>
<td>3</td>
<td>5</td>
<td>65.26</td>
</tr>
<tr>
<td>commission</td>
<td>7</td>
<td>11231</td>
<td>59.81</td>
</tr>
<tr>
<td>karlsruhe</td>
<td>3</td>
<td>22</td>
<td>57.50</td>
</tr>
<tr>
<td>seize</td>
<td>4</td>
<td>2342</td>
<td>42.17</td>
</tr>
<tr>
<td>plutonium</td>
<td>3</td>
<td>437</td>
<td>39.94</td>
</tr>
<tr>
<td>suitcase</td>
<td>3</td>
<td>752</td>
<td>36.69</td>
</tr>
<tr>
<td>german</td>
<td>4</td>
<td>12738</td>
<td>28.69</td>
</tr>
<tr>
<td>material</td>
<td>4</td>
<td>18418</td>
<td>25.78</td>
</tr>
<tr>
<td>seizure</td>
<td>2</td>
<td>443</td>
<td>24.95</td>
</tr>
</tbody>
</table>

Text Normalisation

- Linguistic pre-processing = normalisation of the text
  - **Lemmatisation** (base-form reduction of words) and lower-casing: Transporting → transport
  - Mark-up of multi-word expressions
    - plant → ‘green_plant’ vs. ‘power_plant’
  - **Stop word lists** to avoid words that are not content-bearing
    - general: are, they, having, in_spite_of, interesting
    - domain-specific: question, answer, commission, article
Training: Produce Associate Lists

- Using a large collection of manually indexed documents (training corpus)
- For each descriptor $D_i$, take all documents indexed with $D_i$
- Identify the statistically salient words in each of these texts
- Join these lists of statistically salient words and take the most frequently occurring words as associates. E.g. descriptor RADIOACTIVE MATERIALS

**Result of Training:** Weighed associate lists for all descriptors

### Associate List: RADIOACTIVE MATERIALS

- **rutherford**
- **lithium** 33.0805724759999
- **uranium** 32.506708225322
- **titium** 22.062645151943048
- **nuclear_material** 13.79990098627
- **radioactive_material** 7.8497067551556
- **plutonium** 6.72955494190221
- **radioactive_substance** 6.43422856440347
- **nuclear** 5.651612117597
- **Undine_uta_bloch_von_blohnitz** 5.53276596694963
- **radioactive** 4.893993000302335
- **nuclia_sharn** 4.04705620359469
- **radion** 4.0335435560448
- **mox** 3.56541647221
- **uranium** 3.33654820260962
- **illegal_traffic** 3.0307838955364
Automating Eurovoc Descriptor Assignment

Associate List: **FISHERY MANAGEMENT**

- fishery_resource
- fishery
- fishing
- fish
- ichthyology
- fishing_activity
- fishing_aspect
- fishery
- fishing_location
- fishing_management
- fishery_management
- fishery_resource
- fish
- fishing
- fish
- ichthyology
- fishing_aspect
- fishery
- fishing_location
- fishing_management
- fishery_management

Associate List: **MAURITANIA**

- Mauritanian
- Republic_of_Mauritania
- hole_and_line_fishing
- fishery
- fishing
- fishing_management
- fish
- ichthyology
- fishing_aspect
- fishery
- fishing_location
- fishing_management
- fishery_management

EC - Joint Research Centre - IPSC - CSCF  ---  Ralf Steinberger  9
Assignment Phase

- Pre-process new document (lemmatise, multi-word mark-up)
- Produce lemma frequency list (excluding stop words)
- Calculate similarity between lemma frequency list and descriptor associate lists, using statistical formulae

\[ \text{COSINE}(d, j) = \frac{\sum_{t \in d} \sum_{j \in \text{TFIDF}(t)} \text{TFIDF}(t, d) \cdot \text{TFIDF}(t, j)}{\sqrt{\sum_{t \in d} \sum_{j \in \text{TFIDF}(t)} \text{TFIDF}(t, d)^2} \cdot \sqrt{\sum_{t \in d} \sum_{j \in \text{TFIDF}(t)} \text{TFIDF}(t, j)^2}} \]

Assignment Result (Example)

Title: Legislative resolution embodying Parliament's opinion on the proposal for a Council Regulation amending Regulation No 2847/93 establishing a control system applicable to the common fisheries policy (COM(95)0256 - C4-0272/95 - 95/ 0146(CNS)) (Consultation procedure)
Evaluation of the Assignment

- Separate training and test sets
  - Train on training document set
  - Assign to test document set (ca. 600 documents)

- Compare automatic assignment with previous manual assignment
  - For each rank, calculate
    - Precision (correct assignments divided by all assignment up to this rank)
    - Recall (correct assignments up to this rank divided by no. of man. assigned descr.)

Difficulty of evaluation

- BTs, NTs and RTs have to be considered.
- Number of manually assigned descriptors is small (average 5.6 per text)
- Many other descriptors are also correct.
- (Human) indexing specialists differ in their descriptor assignment (20-80% overlap).

- Additional evaluation of automatically assigned descriptors by human indexer ('manual evaluation'). This provides information:
  - regarding appropriateness of descriptors assigned automatically, but not manually
  - on assignment overlap between two human indexers
Automating Eurovoc Descriptor Assignment

Manual Evaluation of the Assignment

EUROPEAN COMMUNITY
THE EUROPEAN COMMUNITY, hereinafter referred to as the “Community”, and THE KINGDOM OF MOROCCO, hereinafter referred to as “Morocco”, have agreed to cooperate in the management of the fishery sector between the two parties, with a view to the conclusion of an agreement on the exchange of information (the “Cooperation Agreement”).

AWARE of the role played by the sea fisheries sector and related industries in the economic and social development of Morocco and the Community and taking account of both parties’ determination to modernise and rationalise their fishing fleets, each for its part,

RECALLING that the Community and Morocco are signatories to the United Nations Convention on the Law of the Sea of 1982 and that, in accordance with that Convention, Morocco has established an Exclusive Economic Zone...

Eurovoc Assignment – Discussion (1) (Good Result)
Eurovoc Assignment – Discussion (2) (Bad Result)

While the Commission considers that the assessment study provides sufficient assurance that the estimated growth figures are reliable, HAS ADOPTED THIS DECISION:

**Article 1:** The seat capacity restrictions imposed on Air Lingus by Decision 54/119/EC of December 1993 in respect of the rules between Ireland and the United Kingdom and Dublin and London Heathrow, are adjusted, according to Article 5(1) of the Decision to reflect the corresponding market growth in 1994 and 1995, as follows:

- **B-LUK:** 1994: 3 570 765 seats; 1995: 3 043 714 seats.

**Article 2:** This Decision is addressed to Ireland. Done at Brussels, 30 November 1994.

For the Commission:

Marc Van Pronk, Member of the Commission

---

Manual Evaluation – Overview

- **162** documents evaluated.
- Second evaluator reviewed previous manual assignment blindly.
- **Task:**
  - evaluate top ten automatic suggestions (rank 10) and
  - add missing descriptors where necessary
  - Distinguish Good, Bad, BT/NT, ?, S.
- **Averages:**
  - 7.5 correct descriptors per text
  - + 0.65 descriptors (BT or NT)
  - Total: 8.15 (incl. BT and NT)
- **Evaluation of previous manual assignment:**
  - 74% judged as ‘Good’
  - 4% judged as ‘BT’ or ‘NT’
  - Total: 78% agreement = benchmark for automatic assignment
Manual Evaluation - Results

Correct descriptors compared to benchmark of manual assignment (78% G + BT + NT):

\[
\frac{67}{78} = 86\%
\]

Open questions:
- What about the 33% incorrect ones (B + S + ?)
- Where to find the 37% missing descriptors?
- How many descriptors to present?
- How to avoid BT-NT co-occurrence?

Some Difficult Cases (1)

Some descriptors are irrelevant, although many good associates were found, e.g. plutonium sample text

Automatically assigned descriptors:
- plutonium
- import
- illicit trade
- Federal Republic of Germany
- EAEC Joint Research Centre
- airport
- fraud

Manually assigned descriptors:
- plutonium
- import
- illicit trade
- Federal Republic of Germany
- EAEC Joint Research Centre
- airport
- fraud
Some Difficult Cases (2)

- Wrong assignment of semantically related descriptors with similar associate lists:

  RADIOACTIVE WASTE vs. TRANSPORT OF DANGEROUS GOODS

<table>
<thead>
<tr>
<th>Lemmas</th>
<th>Lemmas</th>
</tr>
</thead>
<tbody>
<tr>
<td>radioactive_waste</td>
<td>radioactive_goods</td>
</tr>
<tr>
<td>nuclear_waste</td>
<td>hazardous</td>
</tr>
<tr>
<td>field</td>
<td>key</td>
</tr>
<tr>
<td>nuclear</td>
<td>damage</td>
</tr>
<tr>
<td>process</td>
<td>nuclear</td>
</tr>
<tr>
<td>source</td>
<td>fuel</td>
</tr>
<tr>
<td>waste</td>
<td>shipment</td>
</tr>
<tr>
<td>element</td>
<td>radioactive_waste</td>
</tr>
<tr>
<td>nuclear_power_station</td>
<td>shipment</td>
</tr>
<tr>
<td>radioactive_material</td>
<td>radioactive_material</td>
</tr>
</tbody>
</table>

Current State of our Work (1)

- System is currently optimised for **English and Spanish**
- System is trained for another eight languages without pre-processing:
  De, It, Pt, Ni, Da, Sv, Fi; Fr with using stop words only

![Graph showing performance with and without linguistic pre-processing]
Automating Eurovoc Descriptor Assignment

Current State of our Work (2)

- We work with Eurovoc version 3.1
- In English, associate lists exist for 2565 descriptors
- Many Eurovoc descriptors are not used. In the over 25000 training texts of half a page or more:
  - 35% have never been used!
  - 9% were only used once
  - 23% were only used 2-8 times
- We cannot assign descriptors without training material.

Our Offer and Request

- We would like our software tool to be actively used.
  - Fully automatic assignment for texts that would not be indexed at all.
  - As an interactive tool to support human indexers.
- We need feedback on our work.
- We need more manual evaluation of automatic assignment results.
- Any volunteers? Please contact me at:
  Ralf.Steinberger@jrc.it
  Tel: +39 – 0332 786271
  http://www.jrc.it/langtech