



Cross-lingual Keyword Assignment

XVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)

XVII Conference of the Spanish Society for Natural Language Processing

Jaén, Spain, 12 - 14 September 2001

Ralf Steinberger

European Commission – Joint Research Centre (JRC)

Institute for the Protection and Security of the Citizen (IPSC)

Cyber-security and New Technologies for Combating Fraud

Anti-fraud Information Management Sector (AIM)

<http://www.jrc.it/langtech>



Agenda



- Introduction: Who we are and what we do
- Monolingual keyword assignment (indexing)
- Controlled vocabulary indexing using the multilingual Eurovoc thesaurus
 - Eurovoc thesaurus
 - Statistical Assignment Method
 - Training phase
 - Assignment phase - different algorithms
 - Discussion and evaluation of the results
 - Challenges and difficulties
 - Related work
- Outlook




Institute for the Protection
and Security of the Citizen

Who we are and what we do



JOINT
RESEARCH
CENTRE
EUROPEAN COMMISSION


- The **JRC** is a Directorate General (DG) of the European Commission (www.jrc.it)
- Employs ca. 2500 people in 8 institutes in 5 locations (I, E, D, NL, B)
- **Ispra (I)**: ca. 1800 people
- **Mission**: Carry out scientific research and provide scientific services for DGs and for Member States in a wide range of subjects



Institute for the Protection
and Security of the Citizen

Language Technology in the AIM Sector

<http://www.jrc.it/langtech>



JOINT
RESEARCH
CENTRE
EUROPEAN COMMISSION

- Goal: put together a modular system with three main components:
 - **Retrieval** of potentially relevant texts in a variety of languages, using agent technology
 - **Extraction** of a variety of information aspects from these texts; when possible: language-independent representation of the contents
 - recognise **key words**, subject domains and language of texts, references to geographical places, to people, to products, etc.
 - Calculation of the similarity of documents; clustering and classification of documents
 - **Visualisation** of the contents
 - of individual documents in *document profiles*
 - of whole text collections in *document maps* (Steinberger et al., 2000; Hagman et al. 2000)

Motivation – Methods Used


- **Motivation:** give cross-language access to information 'hidden' in large multilingual document collections
(“fight the information overflow”, “overcome the language barrier”)
- For political and practical reasons: all 11 official EU-languages

- Small team: Johan Hagman (johan.hagman@jrc.it)
Bruno Pouliquen (bruno.pouliquen@jrc.it)
Ralf Steinberger (ralf.steinberger@jrc.it)


- Usage of mainly **statistical methods**
 - less labour-intensive
 - developed methods can easily be adapted to further languages

Monolingual Keyword Identification

- **Statistical approach**
- Minimal linguistic input
 - lemmatisation (base form reduction of words)
 - extensive stop word lists
 - mark-up of multi-word terms (MWU)
- Comparison of the text lemma (word) frequency list with the lemma frequency list of a reference corpus (e.g. BNC, several years of newspaper text)
- comparison of the frequency tables using the *log-likelihood* (or *chi-square*) tests
- a list of keywords and their keyness
(weight; relevance for the document contents)




Monolingual Keyword Identification Example




Document: *Question to the European Parliament regarding a case of smuggled plutonium, seized at Munich airport. Plutonium was analysed by JRC-TUI in Karlsruhe*

KEYWORD	KEYNESS	KEYWORD	KEYNESS
TUI (3/5)	65.31	PSE	17.06
Commission (7/9484)	62.27	Schulz	16.46
Karlsruhe (3/22)	57.55	Euratom	15.99
seizure	55.84	Joint	14.11
OJ	42.21	Germany	12.83
plutonium	39.78	authority	11.79
suitcase	38.44	directorate	11.78
German	29.49	answer	11.58
material	28.51	question	11.56
Munich	23.60	safeguards	11.05
Breyer	22.52	sensational	11.04
airport	17.80	alert	10.98



Advantages and Limitations of this Method




- **Advantage:**
 - To extend to other languages, only lemmatiser / stemmer and reference corpus are needed
- **Limitations of this keyword identification procedure**
 - No compounds apart from the closed list of MWUs (z.B. 'power plant')
 - Monolinguality (multi-monolinguality)
 - Lack of abstraction and consistency ('bread' vs. 'toast' vs. 'bakery products')
- Professional organisations use people to assign controlled vocabulary keywords from a multilingual thesaurus (e.g. Eurovoc)
- ➔ We want to assign Eurovoc descriptors automatically

Eurovoc Thesaurus

- Developed by the European Parliament (EP) and the EC's Publications Office (OPOCE), together with several national organisations
- Controlled Vocabulary
- Multilingual (exists in all 11 official EU languages) !
- We have access to large amounts of training material (manually indexed texts)
- Hierarchically organised into max. 8 levels
 - 21 fields (*politics; law; economics; social questions; environment; industry; geography; energy; agri-foodstuffs; agriculture, forestry and fisheries; international organisations; etc.*)
 - 127 micro-thesauri
 - 5933 descriptors
 - 5877 reciprocal relations (BT, NT), 2730 reciprocal associations (RT)
- Challenge: Descriptor terms like 'DEMOGRAPHY AND POPULATION' or 'CONSTRUCTION AND TOWN PLANNING' are unlikely to occur as such in a text


Assignment of Eurovoc Descriptors

- **Training phase**
 - Produce, for each descriptor, lists of general language lemmas (words) which are 'associated' with this descriptor (*associates*) by
 - compiling *metatexts* containing all documents which were manually indexed with a descriptor
 - using the monolingual keyword assignment tool to identify the most pertinent words (keywords, associates) of this metatext, plus their *weight* (keyness, association strength)
 - **Assignment phase**
 - Compare the lemma frequency list of a new text with all descriptor associate lists
 - Use a statistical algorithm to calculate which descriptor list is most relevant to the text
- ➔ **Result:** a ranked list of the most suitable descriptors for this text




Examples for 'Associates'

'Fishery_Management' & 'Democracy'




fishery	2751.07	human	1007.52
fish	1743.80	right	939.07
stock	1653.37	democracy	892.03
fishing	1191.11	operation	450.15
conservation	826.47	democratic	408.99
management	731.24	ombudsman	359.25
vessel	720.05	freedom	270.69
flag	533.36	fundamental	245.70
organization	525.05	cuba	211.33
agreement	493.99	principle	192.35
migratory	424.20	russia	185.05
subregional	422.25	consolidate	184.68
catch	390.41	political	182.20
mediterranean	323.22	cooperation	177.99
sea	320.55	respect	174.74
highly	312.76	country	144.50
session	263.72	situation	130.41
resource	258.71	turkey	129.87
arrangement	252.56	general	127.28
fly	250.37	finance	110.42
fleet	214.19	headquarters	103.17
gfcml	202.66	relation	100.35
fisherman	198.93	election	98.75
regulation	181.7	subsidiarity	96.82
...		...	



Calculation of the Descriptor Score

Document title: "Seizure of Plutonium at Munich Airport"



Initial, intuitive algorithm: Multiply the text lemma frequency with the log of the keyness of each descriptor and add the result to the score of the descriptor; divide the final score by the text length

Score	Descriptor	Associates and their weight
47	nuclear safety	research (2 * 4) + euratom (1 * 6) + reply (1 * 3) + commission (7 * 3) + source (1 * 4) + plutonium (3 * 6) + nuclear (1 * 8) + schulz (1 * 3) + question (2 * 4) + breyer (1 * 3) + safeguard (1 * 4) + material (4 * 6) + munich (2 * 4)
46	radioactive waste	euratom (1 * 4) + aware (1 * 3) + commission (7 * 4) + incident (1 * 3) + german (4 * 3) + plutonium (3 * 5) + nuclear (1 * 7) + question (2 * 5) + germany (2 * 3) + element (1 * 3) + material (4 * 4)
43	plutonium	euratom (1 * 4) + reply (1 * 4) + seizure (2 * 3) + commission (7 * 3) + plutonium (3 * 6) + nuclear (1 * 6) + schulz (1 * 3) + question (2 * 4) + element (1 * 3) + breyer (1 * 3) + material (4 * 5) + munich (2 * 3)
...

TF.IDF (Salton, G. & C. Buckley, 1988: *Term Weighting Approaches in automatic text retrieval*. Information Processing and Management, vol. 1, 24, N° 5, pp. 513-523)

Cosine (Salton, G, 1989: *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*. Reading, Mass., Addison-Wesley)

Okapi (Robertson, S. E., S. Walker, M. Hancock-Beaulieu & M. Gatford, 1994: *Okapi at TREC-3*, Text Retrieval Conference TREC-3, U.S. National Institute of Standards and Technology, Gaithersburg, USA. NIST Special Publication 500-225, pp. 109-126)


Expected Mutual Information Measure (EMI) (Church K. & P. Hanks, 1989: *Word association norms, mutual information, and lexicography*. In ACL Proceedings, 27th Annual Meeting, Vancouver, pp. 76-83)

Home-grown: e.g. Descriptor Score \pm Sum of all (text lemma count * absolute frequency of associate / total frequency of lemma in training collection)


Mixed algorithms

...

- Keyword assignment is an abstract, highly conceptual task
- There are no clear rules which say that a certain descriptor is suitable or not
- Assignment results differ from one person to the next, sometimes even from one day to the next for the same person, and in any case they differ over time because the historical view changes.
- 80%, 60%, 30% overlapping results between people in different experiments
- Judging the automatic results comparing them to the manually assigned descriptors is an approximation.
- The manual results are not an absolute criterion for the assignment quality.




Automatically Assigned Descriptors




Top 23 English Eurovoc descriptors and their score assigned automatically to the Spanish policy document *Postura de la Unión Europea frente al descubrimiento del contrabando de plutonio (Attitude of the European Union towards the discovery of plutonium smuggling)* (383 words long)

Score	Descriptor (En)	Score	Descriptor (En)
97	<u>NUCLEAR SAFETY</u>	16	<u>EUROPOL</u>
62	<u>NUCLEAR NON-PROLIFERATION</u>	14	NUCLEAR ACCIDENT
43	<u>NUCLEAR FUEL</u>	14	BUDGETARY DISCHARGE
42	<i>NUCLEAR POWER STATION</i>	13	<i>UKRAINE</i>
38	<i>NUCLEAR TEST</i>	13	<u>CIS COUNTRIES</u>
34	<u>IAEA</u>	12	TRANSPORT OF DANGEROUS GOODS
32	<i>RADIOACTIVE WASTE</i>	12	RESEARCH AND DEVELOPMENT
29	<u>RADIOACTIVE MATERIALS</u>	12	EC-GENERAL BUDGET
28	<i>NUCLEAR ENERGY</i>	11	<u>POLICE COOPERATION</u>
25	<u>ILLICIT TRADE</u>		...
21	<i>DECOMMISSIONING OF POWER STATIONS</i>		<u>Underlined:</u> manually assigned descriptors
18	<i>EABC</i>		<i>Italics:</i> further 'reasonable' descriptors
18	<u>ORGANIZED CRIME</u>		Normal: wrong, but semantically related
17	<i>CIS</i>		Strike through: wrong descriptors, semantically not related




Automatically Assigned Descriptors (Currently Best Results)




Top 23 English Eurovoc descriptors and their score assigned automatically to the Spanish policy document *Postura de la Unión Europea frente al descubrimiento del contrabando de plutonio (Attitude of the European Union towards the discovery of plutonium smuggling)* (383 words long; 11 manually assigned descriptors)


Score	Descriptor (En)	Score	Descriptor (En)
84	<u>IAEA</u>	39	<i>NUCLEAR POWER STATION</i>
82	<u>PLUTONIUM</u> (NT TO <u>RADIOACTIVE MATERIALS</u>)	37	<i>DECOMMISSIONING OF POWER STATIONS</i>
79	<u>NUCLEAR FUEL</u>	37	<i>FUEL REPROCESSING</i> (RT TO <u>NUCLEAR FUEL</u>)
72	<u>RADIOACTIVE MATERIALS</u>	37	<i>RADIOACTIVE WASTE</i>
71	<u>ILLICIT TRADE</u>	37	<i>NUCLEAR ENERGY</i>
67	<u>NUCLEAR SAFETY</u>	35	<u>EUROPOL</u> (Rank 20)
65	<u>NUCLEAR NON-PROLIFERATION</u>	35	<i>EABC</i>
59	<u>POLICE COOPERATION</u>	35	TECHNOLOGICAL CHANGE
55	<u>CIS COUNTRIES</u>	34	<i>CIS</i>
44	<i>PRAEFUL USE OF ENERGY</i> (RT TO <u>NUCLEAR SAFETY</u>)...		...
43	<i>ACTION PROGRAMME</i>		<u>Underlined:</u> manually assigned descriptors
41	<u>ORGANISED CRIME</u> (Rank 12)		<i>Italics:</i> further 'reasonable' descriptors
40	<i>COMMUNITY PROGRAMME</i>		Normal: wrong, but semantically related
40	<i>NUCLEAR TEST</i>		Strike through: wrong descriptors, semantically not related




Assignment Results on Training Collection (Spanish)



MaxRank	First Experiments 28 June 2001		Latest Experiments 4 September 2001	
	Recall CT	Precision CT	Recall CT	Precision CT
1	9%	62%	13%	89%
2	15%	54%	23%	81%
3	21%	47%	32%	74%
5	28%	39%	45%	61%
7	34%	33%	53%	52%
10	40%	27%	62%	42%
15	47%	21%	70%	32%
20	52%	17%	75%	26%
25	56%	15%	79%	22%
30	58%	13%	82%	19%
50	66%	9%	88%	12%
100	73%	5%	93%	7%
Formula:	TF * log (keyness)		0.61 Cosine + 0.21 Okapi + 0.18 SumTfidf	
Descriptors:	2870		3643	
Precision with arbitrary distribution:	0.24%		0.19%	
Training material / descriptor	> 10KB		> 2KB	
Aver. No. of manually assigned descr.	6.91		6.91	
Documents in test collection	3743		3955	



Formula of the Currently Best Experiments (Bruno Pouliquen)



$$\Phi = 0.61 \frac{COSINE}{\max(COSINE)} + 0.21 \frac{Okapi}{\max(Okapi)} + 0.18 \frac{SumTfidf}{\max(SumTfidf)}$$

$$TFIDF_{l,d} = TF_{l,d} \cdot ((\log_2 \frac{N}{DF_l}) + 1)$$

$$COSINE(d,t) = \frac{\sum_{l \in d \cap t} TFIDF_{l,d} \cdot TFIDF_{l,t}}{\sqrt{(\sum_{l \in d} TFIDF_{l,d}^2) \cdot (\sum_{l \in t} TFIDF_{l,t}^2)}}$$

$$SumTfidf(d,t) = \sum_{l \in d \cap t} TFIDF_{l,d} \cdot TFIDF_{l,t}$$

TFIDF = Term Frequency,
Inverse Document Frequency

l = lemma,
d = Eurovoc descriptor
|d| = number of associates in descriptor (size)
M = average size of descriptors
t = new text
N = number of descriptors used
DF = document frequency (n° of descriptors for which the lemma is an associate)

$$Okapi_{t,d} = \sum_{l \in t \cap d} \log\left(\frac{N - DF_l}{DF_l}\right) \frac{TF_{l,d}}{TF_{l,d} + \frac{|d|}{M}}$$

Challenges and Difficulties

- **There is not enough training material for all descriptors**
Training material for < 3700 Es, En and De descriptors > 2KB (½ page)
- **Training material with very different text sizes** ranging from short titles to 20-page texts
- **The training material is very much biased by the interests of the EP**
 - e.g. many 'associates' of the descriptor MAURITANIA pertain to the semantic field 'fishery'
- Some descriptors were used thousands of times, others never
 - ➔ *very different lengths of 'associate' lists*
 - ➔ *very different assignment likelihood*

Ferber's Approach (1997)

Reginald Ferber (1997). *Automated Indexing with Thesaurus Descriptors: A co-occurrence Based Approach to Multilingual Retrieval*. In Peters Carol & Thanos Costantino (eds.) *Research and Advanced Technology for Digital Libraries. 1st European Conference (ECDL'97)*. Springer Lecture Notes, Berlin, pp. 232-255.

- Goal: indexing for cross-lingual document retrieval
- Corpus: 80.000 manually indexed *titles* of publications
 - very homogeneous corpus
- OECD thesaurus, ca. 4000 descriptors in 4 languages
- uses the absolute word frequency (instead of the 'keyness')
- calculates the descriptor-word association with a variation of the *Expected Mutual Information Measure* (EMIM): $p(i\&j)/p(i)^x \cdot p(j)^y$
(Church & Hanks, 1989: *Word association norms, mutual information, and lexicography*)
- no consideration of the hierarchical structure (BTs and NTs) and of RTs
- good results

Outlook - To Do

- Improve performance
 - Improve performance by identifying optimal algorithm
 - Test and tune with parallel texts (texts and their translations)
 - Test on documents that are *not* part of the training corpus

- Apply to real world scenarios
 - Apply to further languages (currently trained for En, Es and De)
 - Incorporate with other tools such as crawler and extraction software