



NewsExplorer – Multilingual News Analysis with Cross-lingual Linking

Machine Learning for Multilingual Information Access

NIPS Workshop -- Whistler, Canada, 10 December 2006

Ralf Steinberger, Bruno Pouliquen, Camelia Ignat

European Commission – Joint Research Centre (JRC) – Italy

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

<http://press.jrc.it/NewsExplorer>

Agenda

- Motivation for cross-lingual document linking
 - EU: need for multilinguality
 - Enhanced Information Extraction by combining information from texts in various languages
 - Demo of NewsExplorer (<http://press.jrc.it/NewsExplorer>)
- Related work
- Overview of our approach
- Description of the components, challenges
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters over time
 - Linking clusters across languages
- Future work and Conclusions

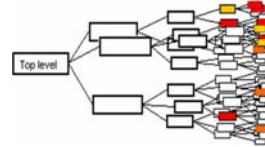
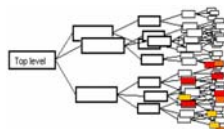
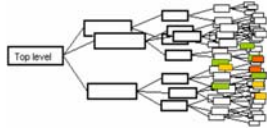
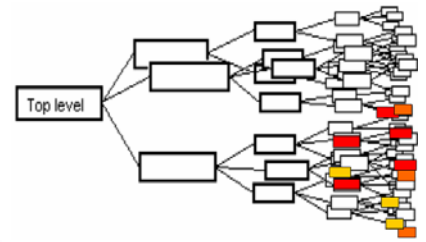
- Usage of Machine Translation → monolingual document similarity (TDT-3, Leek et al. 1999)
- Usage of bilingual dictionaries → monolingual document similarity (Wactlar 1999; Mathieu et al. 2004)
- Automatically produce bilingual lexical space for bilingual document representation and document similarity calculation, e.g.
 - bilingual *Lexical Semantic Analysis* (LSI, Landauer & Littman 1991)
 - *Kernel Canonical Correlation Analysis* (Vinokourov et al., 2002)
- + Achieved results are relatively good
- Bilingual approach is restricted to a few languages:

$$\text{Language pairs} = (N^2 - N) / 2 \quad (N = \text{number of languages})$$

- EU: 20 official languages → 190 language pairs (380 language pair directions)!
- **Attractiveness of highly multilingual *interlingua approaches***

- Motivation for cross-lingual document linking
 - EU: need for multilinguality
 - Enhanced Information Extraction by combining information from texts in various languages
 - Demo of NewsExplorer (<http://press.jrc.it/NewsExplorer>)
- Related work
- **Overview of our approach**
- Description of the components, challenges
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters over time
 - Linking clusters across languages
- Future work and Conclusions

- Map documents onto multilingual thesauri, nomenclatures, gazetteers, ... (general, medical, technical, geographical, ...)



- Identify and normalise language-independent text features
 - Numbers; dates; currency expressions; measurement expressions; ...
le treize août 2007, am 13.8.2007, by 13/08/07 → DATE_2007_08_13

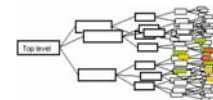
- Names of persons, organisations, ...

Muqtada al-Sadr, Mouqtada Sadr, Mokdata Sadr, Muqtadā aş-Şadr, Муктада Садр
→ ID=[236](#)

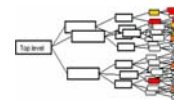
- Calculate the various vector similarities and add up
 - For details, see Steinberger et al. (2004, Informatica 28).

- Represent documents by vectors of ~language-independent features

- S1: Multilingual **thesaurus and classification system** Eurovoc



- S2: **Locations** (Latitude-Longitude information)



- S3: Normalised and merged proper **name variants** (persons and organisations)

Keyword	Keyword
100.2470	Jackson
41.5420	newsworld
37.5547	netia
22.0105	mountain
24.3105	boy
24.6051	pink
20.9824	documentary
10.7973	avocet
13.3945	courthouse
15.1224	burg
10.2124	tal
10.0028	spanish
9.6021	california
9.3905	vestib

- S4: Cognates and **numbers**

Keyword	Keyword
100.2470	Jackson
41.5420	newsworld
37.5547	netia
22.0105	mountain
24.3105	boy
24.6051	pink
20.9824	documentary
10.7973	avocet
13.3945	courthouse
15.1224	burg
10.2124	tal
10.0028	spanish
9.6021	california
9.3905	vestib

- CLDS (using cosine) based on these representations

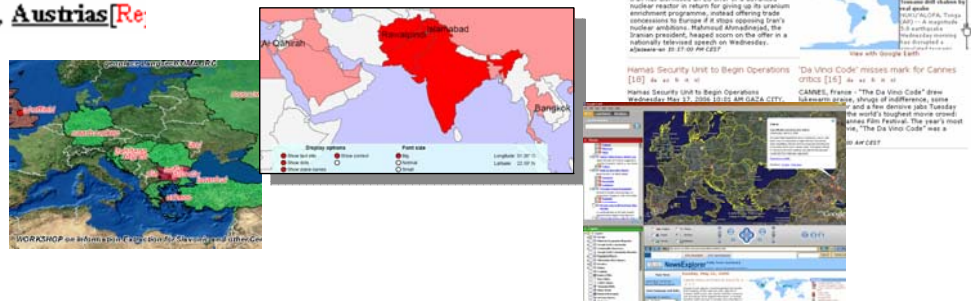
$$CLDS = \alpha \cdot S1 + \beta \cdot S2 + \gamma \cdot S3 + \delta \cdot S4$$

- Motivation for cross-lingual document linking
 - EU: need for multilinguality
 - Enhanced Information Extraction by combining information from texts in various languages
 - Demo of NewsExplorer (<http://press.jrc.it/NewsExplorer>)
- Related work
- Overview of our approach
- **Description of the components, challenges**
 - **IE: Locations**
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters over time
 - Linking clusters across languages
- Future work and Conclusions

regionálním, jako například v **Itali**[Italy], **Sp**
e Francii[France], **Španěls**[Spain]ku, **Rako**
lašších zemích, nebo na základě smíšeného s

→ Latitude / Longitude

i liikmed valitakse otsese ja üldiste valimist
näiteks **Itaalia**[Italy]s, **Ühendkuningrii**[Un
ntsusmaal, **Hispaania**[Spain], **Austrias**[Re



- Place names cannot be recognised by looking for patterns in text (Gey 2000)
- Multilingual gazetteer needed
- Place name recognition via the **lookup** of text words in the gazetteer

1. Places homographic with common words

→ Geo-stopwords

English		French		German	
Place name	Country	Place name	Country	Place name	Country
And	Iran	De	Burkina Faso	Die	France
To	Ghana	Du	Ghana	Den	Ethiopia
Be	India	Un	Russia	Zu	Zaire

2. Places homographic with person names

→ Location only if not part of a person name
e.g. 'Kofi Annan', 'Annan'

Name	City: Country
Tony Blair	Tony: USA
	Blair: Malawi
Kofi Annan	Kofi: Mali
	Annan: Scotland
Javier Solana	Javier: Spain
	Solana: Philippines

3. Completeness of the gazetteer: exonyms, endonyms, ...

'Санкт-Петербург', 'Saint Petersburg', 'Saint Pétersbourg', 'Leningrad', 'Petrograd', ...

4. Inflection

- Romanian: **Parisului** (*of Paris*)
- Estonian: **Londonit** (London),
New Yorgile (New York)
- Arabic: **نوس يرابل** (*the Paris inhabitants*)
[albaRiziu:n]

→ Usage of context-sensitive suffix lists and replacement rules to generate all variants
(Slovene example)

Tony(a|o|u|om|em|m|ju|jem|ja)?s+Blair(a|o|u|om|em|m|ju|jem|ja)

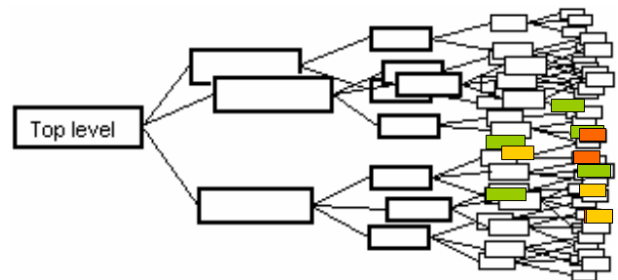
5. Homographic place names

Place name	Number of cities with this name
Aleksandrovka	244
...	
Washington	32
London	18
Berlin	15
Paris	15

- Use size class information
- Use country context
(news source; unambiguous anchors)
- Use kilometric distance
E.g. **“from Warsaw to Brest”**
Brest (France): 2000 km from Warsaw
Brest (Belarus): 200 km from Warsaw

Class	Explanation	Example	Weight
0	country name	Italy	80
1	capital	Rome	80
2	main city	Milan	80
3	province level	Varese	30
4	small city	Sesto Calende	20
5	village	Ispra	10
6	small settlement, hamlet	-	5

For details, see Pouliquen et al. (2006, LREC)



- List of place names found
- Frequency count per country (city, continent, region, ...)
- Frequency can be normalised, using TF.IDF or similar

- Motivation for cross-lingual document linking
 - EU: need for multilinguality
 - Enhanced Information Extraction by combining information from texts in various languages
 - Demo of NewsExplorer (<http://press.jrc.it/NewsExplorer>)
- Related work
- Overview of our approach
- Description of the components, challenges
 - IE: Locations
 - **IE: Person and Organisation names**
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters over time
 - Linking clusters across languages
- Future work and Conclusions

en	death of former Prime Minister Rafik Hariri, blamed by many opposition
es	asesinato del ex primer ministro Rafic al-Hariri, que la oposición atribuyó
fr	l'assassinat de l'ex-dirigeant Rafic Hariri et le départ du chef de la diplom
nl	na de moord op oud-premier Rafiq al-Hariri gingen gisteren bijna een
de	libanesischen Regierungschef Rafik Hariri vor einem Monat wichtige B
sl	danjega libanonskega premiera Rafika Haririja. Libanonska opozicija si
et	möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommipl
ar	اغتيال رئيس الوزراء السابق رفيق الحريري بأياد يهودية وما حدث سابقا
ru	Бывший премьер-министр Ливана Рафик Харири, который

- **Lookup of known names** from database
 - Currently over 560,000 names (excluding spelling variants)
 - ~1000 new names per day
 - Pre-generate morphological variants (Slovene example):
Tony(a|o|u|om|em|m|ju|jem|ja)?\s+**Blair**(a|o|u|om|em|m|ju|jem|ja)
- **Guessing names** using empirically-derived *lexical patterns*
 - **Trigger word(s) + Name Surname**
 - **President, Minister, Head of State, Sir, American**
 - **“death of”, “[0-9]+-year-old”, ...**
 - Known first names (**John, Jean, Hans, Giovanni, Johan, ...**)
 - ...
 - Combinations: **“56-year-old former prime minister Kurmanbek Bakiyev”**

- Recognition of person names, using regular expressions (Slovene example):
 - **kandidat**(a|u|om)?
 - **legend**(a|e|i|o)
 - **milijarder**(ja|ju|jem)?
 - **predsednik**(a|u|om|em)?
 - **predsednic**(a|e|i|o) + **uppercase words**
 - **ministric**(a|e|i|o)
 - **sekretar**(ja|ju|jom|jem)?
 - **diktator**(ja|ju|jem)?
 - **playboy**(a|u|om|em)?

... verskega **voditelja** **Moktade al Sadra** je z notranjim ...
= **Muqtada al-Sadr** (ID=[236](#))

- For all new names found: apply *approximate name matching*
 - Based on sets of letter bigrams and letter trigrams
 - Merge two names if cosine similarity is > 70%
- Collect variants automatically from [Wikipedia](#)
- Cross-script* name matching
(Cyrillic, Greek, Arabic + Farsi, Hindi)
 - Тони Блеър => Tony Blair
 - Ιυιάντ Αλλάουι => Iyad Allaoui
- For details, see Pouliquen et al. (Journal Corela, 12/2005)

name	count	lang(s)
Rafik Hariri	2550	(Eu..sv)
Rafiq Hariri	827	(Eu..pl)
Rafik al-Hariri	494	(de..nl)
Rafic Hariri	315	(Eu..nl)
Rafiq al-Hariri	224	(de..en)
Rafiq Al Hariri	30	en
رفيخ الحاريري	12	ar
Rafik Al Hariri	9	en
Rafik al Hariri	9	de
Rafik el-Hariri	8	de
Rafiq Al-Hariri	6	(da..en)
Rafik Al-Hariri	6	(de..en)
Rafik Hariri Hariri	3	de
Rafik el Hariri	3	de
...

- Frequency list of normalised person names found
(numerical person ID)

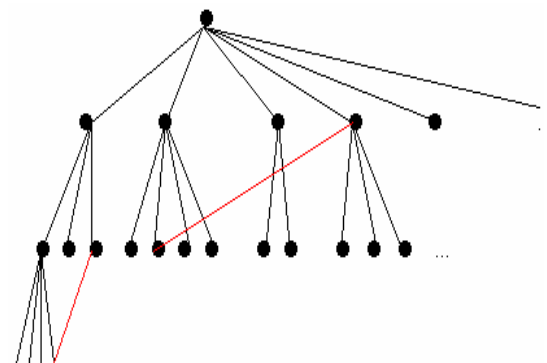
Related People
Alexander Litvinenko (17)
Mario Scaramella (11)
John Reid (8)
Vladimir Putin (8)
Andrei Lugovoi (8)
Anna Politkovskaya (3)
Leonid Nevzlin (3)
Boris Berezovsky (3)
Tony Blair (2)

- Motivation for cross-lingual document linking
 - EU: need for multilinguality
 - Enhanced Information Extraction by combining information from texts in various languages
 - Demo of NewsExplorer (<http://press.jrc.it/NewsExplorer>)
- Related work
- Overview of our approach
- Description of the components, challenges
 - IE: Locations
 - IE: Person and Organisation names
 - **Categorisation into Subject domains (Eurovoc classes)**
 - Clustering of news
 - Linking clusters over time
 - Linking clusters across languages
- Future work and Conclusions

- **Over 6000 classes**
- Covering many different subject domains (wide coverage)
- **Multilingual** (over 20 languages, one-to-one translations)
- Developed by the European Parliament and others
- Actively used **to manually index and retrieve documents** in large collections (fine-grained classification and cataloguing system)
- Freely available for research purposes
- **Hierarchically organised** into up to 8 levels

Relations:

- Broader Terms
- Narrower Terms
- **Related Terms**



- Eurovoc is a conceptual thesaurus

E.g.

- SPORT
- PROTECTION OF MINORITIES
- CONSTRUCTION AND TOWN PLANNING
- RADIOACTIVE MATERIALS

→ *categorisation* vs. *term extraction*

- Large number of classes (~ 6000)
- Very unevenly distributed
- Various text types (heterogeneous training set)
- Multi-label categorisation (both for training and assignment)

- Profile-based, category ranking task
 - Training: Identification of most significant words for each class
 - Assignment: combination of measures to calculate similarity between profiles and new document

- Empirical refinement of parameter settings

- Training:
 - Stop words
 - Lemmatisation
 - Multi-word terms
 - Consider number of classes of each training document
 - Thresholds for training document length and number of training documents per class
 - Methods to determine significant words per document (log-likelihood vs. chi-square, etc.)
 - Choice of reference corpus
 - ...
- Assignment:
 - Selection and combination of similarity measures (cosine, okapi, ...)
 - ...

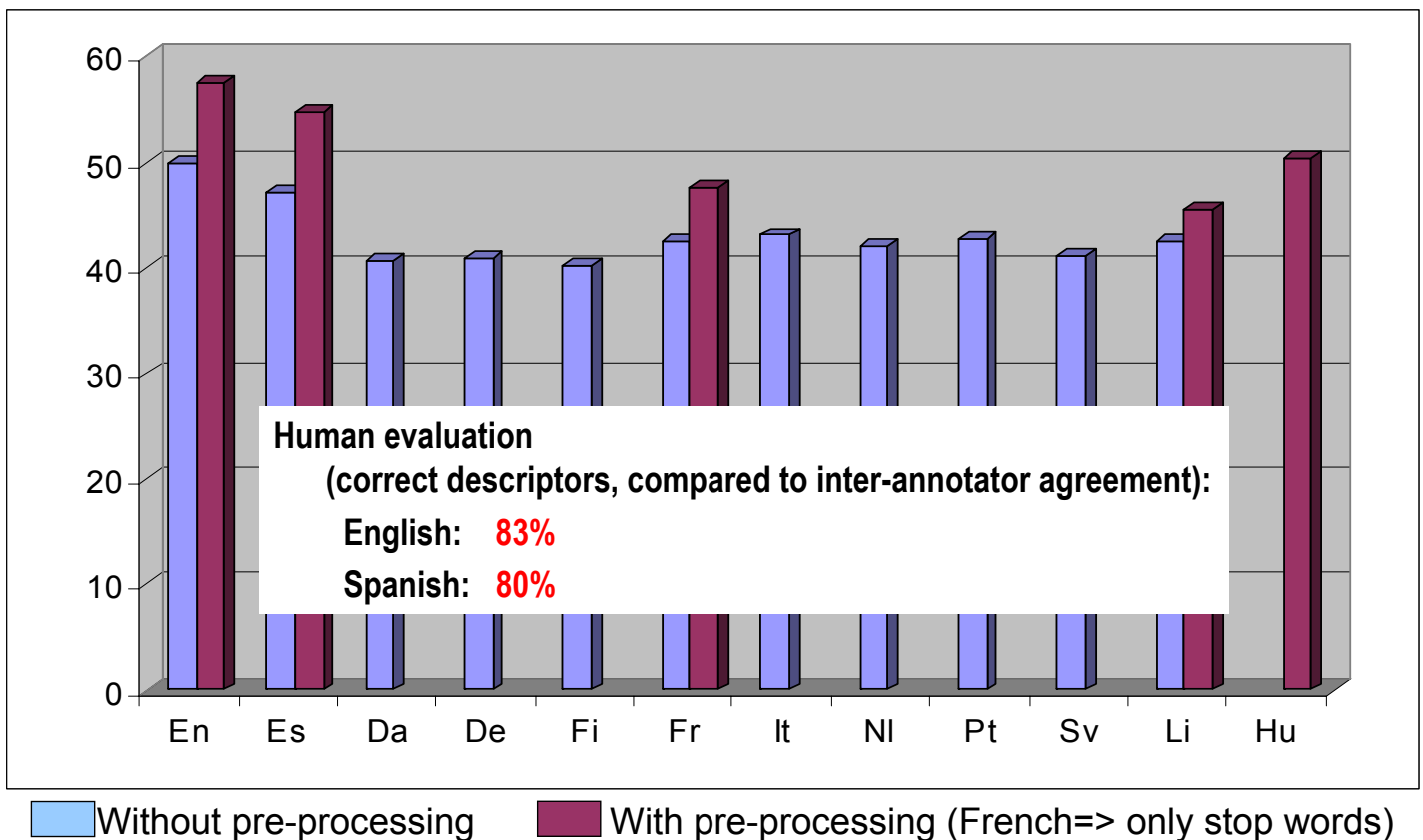
Sample profile: FISHERY MANAGEMENT

Category	Word	Value
fishery-related	fishery_management	14.47113428888
	fishery	14.11111111111
	fish	14.11111111111
	fisheries	14.11111111111
	fishery_management	14.11111111111
	fishery_management	14.11111111111
	fishery_management	14.11111111111
	fishery_management	14.11111111111
	fishery_management	14.11111111111
	fishery_management	14.11111111111
management-related	management	14.11111111111
	management	14.11111111111
	management	14.11111111111
	management	14.11111111111
	management	14.11111111111
	management	14.11111111111
	management	14.11111111111
	management	14.11111111111
	management	14.11111111111
	management	14.11111111111

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)

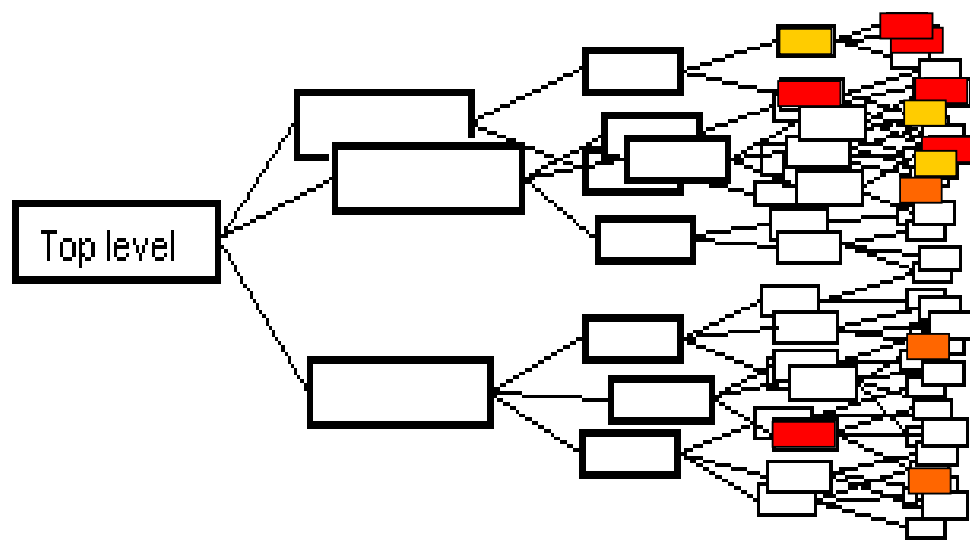
Descriptor ID	Descriptor text	Inverse square Sum Tfidf ²	Cosine	Rank Cosine	Okapi	Rank Okapi	Rank	Prec	Rec
5641040706000000	FISHING CONTROLS [g]	.000144033	0.360	1	95.169	1	1	100	10
5641020000000000	FISHING GROUNDS [nt]	.00243484	0.308	2	65.018	14	2	100	20
5641040200000000	COMMON FISHERIES POLICY [g]	.00018023	0.280	3	62.910	20	3	100	30
5641040100000000	FISHERY MANAGEMENT [nt]	.000207086	0.279	4	79.362	6	4	100	40
5641040700000000	FISHING REGULATIONS [g]	.000197934	0.270	5	79.982	5	5	100	50
5641040704000000	FISHING PERMIT [g]	.00306631	0.261	6	71.577	8	6	100	60
5641040101000000	CONSERVATION OF FISH STOCKS [s]	.000189818	0.253	7	83.982	3	7	85	60
5641040600000000	FISHING AREA [g]	.000182474	0.252	8	84.178	2	8	75	60
5206040100000000	CONSERVATION OF RESOURCES [s]	.000234209	0.251	9	55.311	26	9	66	60
5641050000000000	FISHERY RESOURCES	.000402883	0.232	10	75.046	7	10	60	60
5641040800000000	CATCH OF FISH	.000313101	0.213	11	67.687	9	11	54	60
5641040000000000	FISHERIES POLICY	.00258399	0.203	12	58.416	23	12	50	60
5641040705000000	FISHING LICENCE	.000371136	0.181	13	57.618	25	13	46	60
5641060100000000	FISHING FLEET	.00106478	0.179	14	63.323	19	14	42	60
5641010000000000	FISHING INDUSTRY	.000551953	0.176	15	39.228	43	15	40	60
5641040201000000	EUROPECHE	.000738822	0.176	16	62.240	21	16	37	60
...									

Results of automatic evaluation across languages (F1 per document at rank=6)



- Ranked list of Eurovoc descriptor codes found for each document

Descriptor ID	Cosine
5641040706000000	0.360
5641020000000000	0.308
5641040200000000	0.280
5641040100000000	0.279
5641040700000000	0.270
5641040704000000	0.261
5641040101000000	0.253
5641040600000000	0.252
5206040100000000	0.251
5641050000000000	0.232
5641040800000000	0.213
5641040000000000	0.203
5641040705000000	0.181
5641060100000000	0.179
5641010000000000	0.176
5641040201000000	0.176
...	...



- **Freely available** for research purposes on our web site: <http://langtech.jrc.it/JRC-Acquis.html>
- For details, see Steinberger et al. (2006, LREC)

- Average of 8.8 Million words per language
- Pair-wise alignment for all 210 language pairs!
- Average of 7600 documents per language
- **Most documents have been Eurovoc-classified manually**

→ useful for:

1. Training of multilingual subject domain classifiers.
2. Creation of multilingual lexical space (LSA, KCCA)
3. Training of automatic systems for Statistical Machine Translation.
4. Producing multilingual lexical or semantic resources such as dictionaries or ontologies.
5. Training and testing multilingual information extraction software.
6. Automatic translation consistency checking.
7. Testing and benchmarking alignment software (sentences, words, etc.), across a larger variety of language pairs.
8. All types of multilingual and cross-lingual research.

- Motivation for cross-lingual document linking
 - EU: need for multilinguality
 - Enhanced Information Extraction by combining information from texts in various languages
 - Demo of NewsExplorer (<http://press.jrc.it/NewsExplorer>)
- Related work
- Overview of our approach
- Description of the components, challenges
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - **Clustering of news**
 - **Linking clusters over time**
 - Linking clusters across languages
- Future work and Conclusions

- Vector of keywords and their keyness using log-likelihood test (Dunning 1993)

“Michael Jackson Jury Reaches Verdicts”

<u>Keyness</u>	<u>Keyword</u>	<u>Keyness</u>	<u>Keyword</u>
109.24	jackson	9.39	verdict
41.54	neverland	7.56	testimony
37.93	santa	6.50	maria
32.61	molestation	4.09	michael
24.51	boy	1.73	reached
24.43	pop	1.68	ap
20.68	documentary	1.05	appeared
18.79	accuser	0.53	child
13.59	courthouse	0.50	trial
11.12	jury	0.45	monday
10.08	ranch	0.26	children
9.60	california	0.09	family

Monday, June 13, 2005

Michael Jackson Jury Reaches Verdicts es de it nl

Jackson, 46, was accused of molesting the then-13-year-old boy and plying him with wine at the pop star's Neverland ranch in 2003. Jackson had befriended the boy, a cancer survivor, and they appeared together when Jackson was interviewed for the documentary "Living With Michael Jackson." ABCnews 13/06/2005 21:54

Jackson cleared of all 10 charges
ananova 13/06/2005 23:36

Timetable Of Events Which Led To Court
skynews 13/06/2005 23:30

Week 2 Of Jackson Deliberations
CBSnews 13/06/2005 12:07

Jackson jurors set for second week
NEWS.comAU 13/06/2005 03:52

Jackson jury into second week
TheAustralian 13/06/2005 18:12

Music's misunderstood superstar
bbc 13/06/2005 23:25

The Extraordinary Life Of A Music Icon
five 13/06/2005 23:32

Jury finds Michael Jackson innocent of all charges
ishkimes 13/06/2005 23:50

Michael Jackson found not guilty on all charges
tv 13/06/2005 23:34

JACKSON NOT GUILTY
skynews 13/06/2005 23:22

Michael Jackson cleared of abuse
bbc 13/06/2005 23:25

Analysis: Testing times ahead
guardian 13/06/2005 23:48

Jackson arrives at court
TheAustralian 13/06/2005 22:27



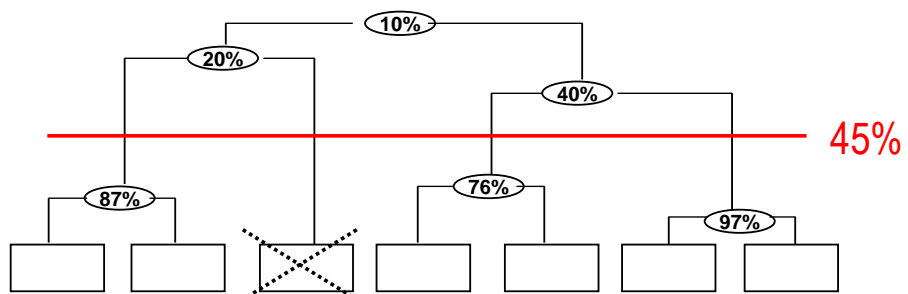
- Aim: show to what extent a text talks about a certain country
- Sum of references to a country, normalised using the log-likelihood test
- Add country score vector to keyword vector

10.4184	*us*
1.5610	*gb*
1.5610	*il*
1.5610	*br*

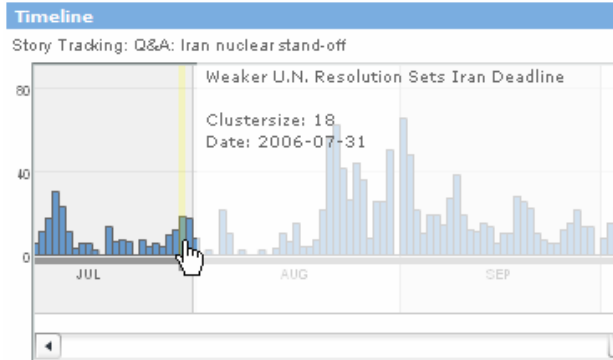
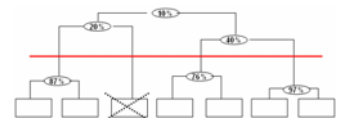
Keyness	Keyword	Keyness	Keyword
109.2478	jackson	7.5620	testimony
41.5450	neverland	6.5014	maria
37.9347	santa	4.0957	michael
32.6105	molestation	1.7368	reached
24.5193	boy	1.6857	ap
24.4351	pop	1.5610	*gb*
20.6824	documentary	1.5610	*il*
18.7973	accuser	1.5610	*br*
13.5945	courthouse	1.0520	appeared
11.1224	jury	0.5384	child
10.4184	*us*	0.5045	trial
10.0838	ranch	0.4502	monday
9.6021	california	0.2647	children
9.3905	verdict	0.0946	family

- Input: Vectors consisting of keywords and country score
- Similarity measure: cosine
- Method: Bottom-up group average unsupervised clustering
- Build the binary hierarchical clustering tree (dendrogram)
 - Retain only "big" nodes in the tree with a high cohesion (empirically refined minimum intra-node similarity: 45%)
- Use the title of the cluster's medoid as the cluster title
- For details, see Pouliquen et al. (CoLing 2004)

Keyness	Keyword
109.2478	jackson
41.5450	neverland
37.9347	santa
32.6105	molestation
24.5193	boy
24.4351	pop
20.6824	documentary
18.7973	accuser
13.5945	courthouse
11.1224	jury
10.4184	*us*
10.0838	ranch
9.6021	california
9.3905	verdict



- Details: Pouliquen et al. (CoLing 2004)
- Evaluation results depending on similarity threshold



Similarity threshold	“Precision”	“Recall”
15%	88%	100%
20%	92%	98%
40%	98%	86%
60%	99%	78%
80%	99%	67%

- Motivation for cross-lingual document linking
 - EU: need for multilinguality
 - Enhanced Information Extraction by combining information from texts in various languages
 - Demo of NewsExplorer (<http://press.jrc.it/NewsExplorer>)
- Related work
- Overview of our approach
- Description of the components, challenges
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters over time
 - **Linking clusters across languages**
- Future work and Conclusions

$$CLDS = \alpha \cdot S1 + \beta \cdot S2 + \gamma \cdot S3 + \delta \cdot S4$$

- Ranked list of Eurovoc classes (40%)
- Country score (30%)
- Names + frequency (20%)
- Monolingual cluster representation *without country score* (10%)

Descriptor ID	Cosine γ
5641040706000000	0.360
5641020000000000	0.308
5641040200000000	0.280
5641040100000000	0.279
5641040700000000	0.270
5641040704000000	0.261
5641040101000000	0.253
5641040600000000	0.252
5206040100000000	0.251
5641050000000000	0.232
5641040000000000	0.213
5641040000000000	0.203
5641040705000000	0.181
5641060100000000	0.179
5641010000000000	0.176
5641040201000000	0.176

+ **10.4184** *us*
1.5610 *gb*
1.5610 *il*
1.5610 *br*

Related People
Kim Jong Il (10)
Stephen Hadley (9)
Shinzo Abe (5)
Junichiro Koizumi (5)
Condoleezza Rice (5)
Tony Snow (5)
Donald Rumsfeld (3)
John Bolton (3)
Christopher Hill (3)

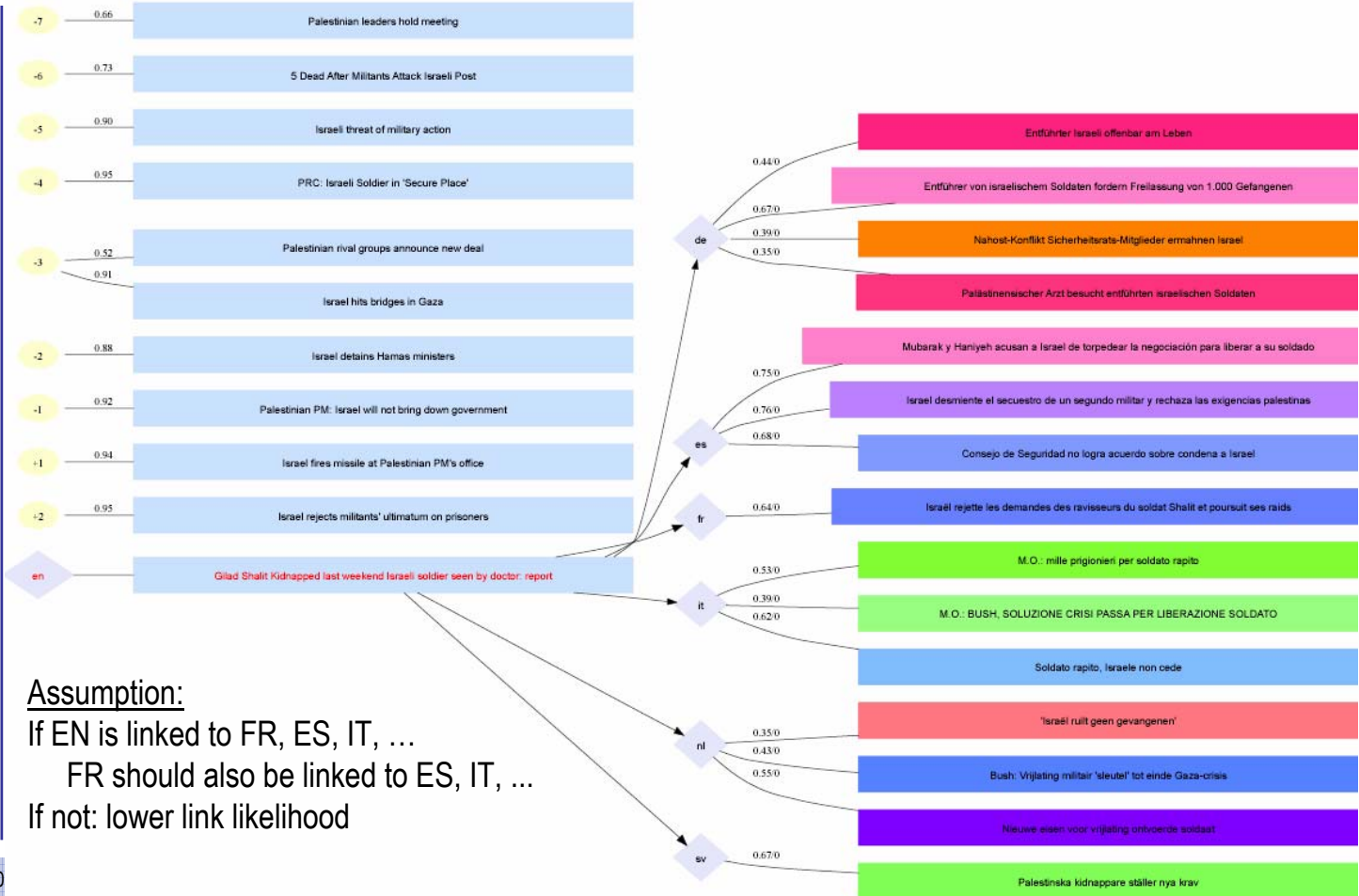
Keyness	Keyword
109.2478	jackson
41.5450	neverland
37.9347	santa
32.6105	molestation
24.5193	boy
24.4351	pop
20.6824	documentary
18.7973	accuser
13.5945	courthouse
11.1224	jury
10.0838	ranch
9.6021	california

- Evaluation results depending on similarity threshold
- Ingredients: 40/30/30 (names not yet considered)
- Evaluation for EN → FR and EN → IT (136 EN clusters)

Similarity threshold	EN → FR Precision	EN → FR Recall *	EN → IT Precision	EN → IT Recall *
30%	84%	99%	71%	97%
60%	98%	46%	98%	42%

* Recall at 15% similarity threshold = 100%

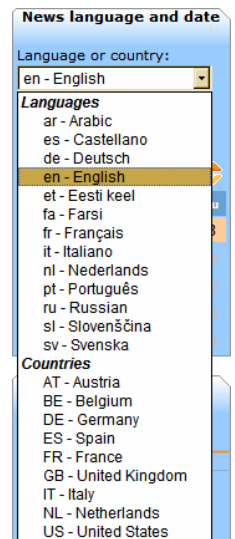
- For details, see Pouliquen et al. (CoLing 2004)



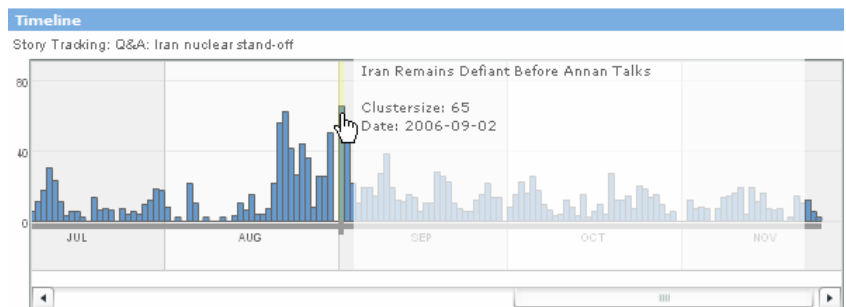
- Motivation for cross-lingual document linking
 - EU: need for multilinguality
 - Enhanced Information Extraction by combining information from texts in various languages
 - Demo of NewsExplorer (<http://press.jrc.it/NewsExplorer>)
- Related work
- Overview of our approach
- Description of the components, challenges
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters over time
 - Linking clusters across languages
- **Future work and Conclusions**

1. Add more languages (analysis and cross-lingual linking)
2. Add more information facets for cross-lingual cluster linking
 - Dates
 - Professions
 - Expressions of measurement
 - Vehicles
 - ...
3. Empirically optimise weighting of ingredients

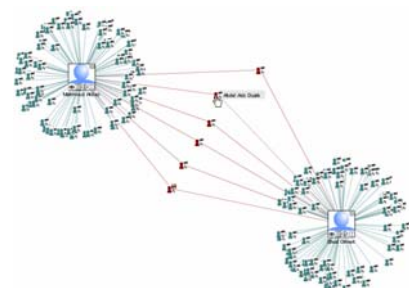
$$CLDS = \alpha \cdot S1 + \beta \cdot S2 + \gamma \cdot S3 + \delta \cdot S4 + \dots$$



4. Link longer-running *stories* to other languages (currently, only individual *clusters* are linked)



5. Categorise persons (governmental, military, religious, civil, ...)
6. Categorise relations between persons
 - Family relation
 - Contact/Meeting information
 - ORG-Membership relation
 - ...



- Cross-lingual linking of documents/clusters via language-independent representations is feasible.
- More information can be extracted when using texts written in several languages.
- Simple means can take you a long way
 - JRC effort to add a new language is 1 - 6 months
 - Simple lexical patterns
 - Simple morphological suffix treatment
 - Clustering and categorisation
 - Language-independent statistics and heuristics.
- Monolingual analysis is sufficient; **no language pair-specific information is needed.**

Alexander Litvinenko

Information about this person was last updated on Monday, December 4, 2006.

Aliases	Key titles and phrases	External resources
Alexander Litvinenko (ru,ru)	russo (it,pt - 181)	
Alexandre Litvinenko (fr)	agent russe (fr - 99)	
Alexander Litwinenko (de)	ruso (es - 115)	
Aleksandr Litvinenko (fi,no)	kriitikeri (de - 21)	
Aleksander Litvinenko (nl,sv)	agent (en,nl - 47)	
Александр Литвиненко (ru)	russo (it - 46)	
Александр Литвиненко (ru)	russe (de,fr - 58)	
Alexandr Litvinenko (it)	agenten (de,sv - 34)	
Alexandre Litvinenko (fr)	agent secret russe (fr - 20)	
Aleksander Lytvynenko (en)	monte di (it - 27)	
Alexander Litvynenko (de)	russo (it - 11)	
Aleksander Litvjenko (hr)	43 ans (fr - 11)	
Александр Валтерович Литвиненко (ru)		
Alexandr Litvinenko (cs)		
Aleksander Litwinenko (pl)		
Alexander Litvinen (fi)		
アレクサンダー・リトビネンコ (ja)		
Alexander Walterowitsch Litwinenko (de)		




Image obtained automatically from Wikipedia