European Commission
Joint Research Centre

Towards Language-Independent News Summarization

Josef Steinberger
Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Marco Turchi, Vanni Zavarella

http://www.jrc.ec.europa.eu
http://ipsc.jrc.ec.europa.eu/

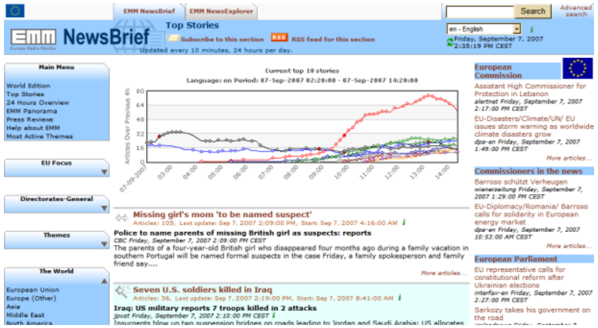
Outline

- **Motivation** – summaries of news clusters in Europe Media Monitor (EMM)
- **Summarization approach**
  - Basic approach based on latent semantic analysis (LSA) – TAC'08
  - Adding semantic information about entities – TAC'09
  - Aspect capturing
    - event extraction + semantic class learning – TAC'10
  - Temporal analysis
  - Sentence compression and paraphrasing by term sequence selection and sentence reconstruction inspired by MT techniques – TAC'11
- Results in the **guided summarization task**
- Our prior work on multilingual evaluation
- Results in the **multilingual summarization task**

Europe Media Monitor

- **EMM news gathering engine**
  - Monitors ~ 3,000 news sources
  - Gathers about 100,000 news articles per day
  - In **>50 languages**
  - Visits some sites every 5 minutes
  - Extracts text from the web page
  - Converts text into Unicode-encoded RSS
  - Feeds the news into publicly accessible media monitoring systems

http://emm.newsbrief.eu/overview.html



Top Stories

Subscribe to this section RSS RSS feed for this section

Updated every 10 minutes, 24 hours per day.

Friday, September 7, 2007
2:35:19 PM CEST

Main Menu

World Edition
Top Stories
24 Hours Overview
EMM Panorama
Press Reviews
Help about EMM
Most Active Themes

The World

European Union
Europe (Other)
Asia
Middle East
North America

Missing girl's mom 'to be named suspect'
Articles: 105, Last update: Sep 7, 2007 2:09:00 PM, Start: Sep 7, 2007 4:16:00 AM

Police to name parents of missing British girl as suspects: reports
CBC Friday, September 7, 2007 2:09:00 PM CEST
The parents of a four-year-old British girl who disappeared four months ago during a family vacation in southern Portugal will be named formal suspects in the case Friday, a family spokesperson and family friend say....

More articles...

Seven U.S. soldiers killed in Iraq
Articles: 36, Last update: Sep 7, 2007 2:19:00 PM, Start: Sep 7, 2007 8:41:00 AM

Iraq: US military reports 7 troops killed in 2 attacks
jpost Friday, September 7, 2007 2:10:00 PM CEST
Insurgents blow up two suspension bridges on roads leading to Jordan and Saudi Arabia; US allocates

European Commission

Assistant High Commissioner for Protection in Lebanon
alertnet Friday, September 7, 2007 2:17:00 PM CEST

EU-Disasters/Climate/UN/ EU issues storm warning as worldwide climate disasters grow
dpa-en Friday, September 7, 2007 1:49:00 PM CEST

More articles...

Commissioners in the news

Barroso schützt Verheugen
wienerzeitung Friday, September 7, 2007 1:29:00 PM CEST

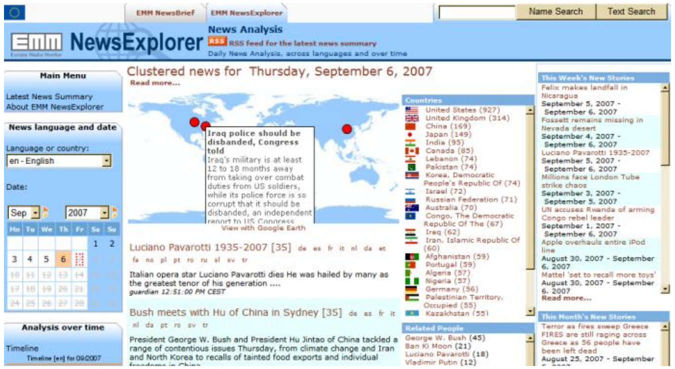
EU-Diplomacy/Romania/ Barroso calls for solidarity in European energy market
dpa-en Friday, September 7, 2007 10:53:00 AM CEST

More articles...

European Parliament

EU representative calls for constitutional reform after Ukrainian elections
interfax-en Friday, September 7, 2007 2:27:00 PM CEST

Sarkozy takes his government on the road



News Analysis

RSS RSS feed for the latest news summary

Daily News Analysis, across languages and over time

Latest News Summary
About EMM NewsExplorer

Language or country:
en - English

Date:

Timeline
Timeline [en] for 09/2007

Clustered news for Thursday, September 6, 2007

Read more...

Iraq police should be disbanded, Congress told
Iraq's military is at least 12 to 18 months away from taking over combat duties from US soldiers, while its police force is so corrupt that it should be disbanded, an independent report to US Congress

Luciano Pavarotti 1935-2007 [35] de es fr it nl da et fa no pl pt ro ru sl sv tr

Italian opera star Luciano Pavarotti dies He was hailed by many as the greatest tenor of his generation ....
guardian 12:51:00 PM CEST

Bush meets with Hu of China in Sydney [35] de es fr it nl da pt ro sv tr

President George W. Bush and President Hu Jintao of China tackled a range of contentious issues Thursday, from climate change and Iran and North Korea to recalls of tainted food exports and individual freedoms in China

Countries

United States (927)
United Kingdom (314)
China (169)
Japan (149)
India (95)
Canada (85)
Lebanon (74)
Pakistan (74)
Korea, Democratic People's Republic Of (74)
Israel (72)
Russian Federation (71)
Australia (70)
Congo, The Democratic Republic Of The (67)
Iraq (62)
Iran, Islamic Republic Of (60)
Afghanistan (59)
Portugal (59)
Algeria (57)
Nigeria (57)
Germany (56)
Palestinian Territory, Occupied (55)
Kazakhstan (55)

Related People

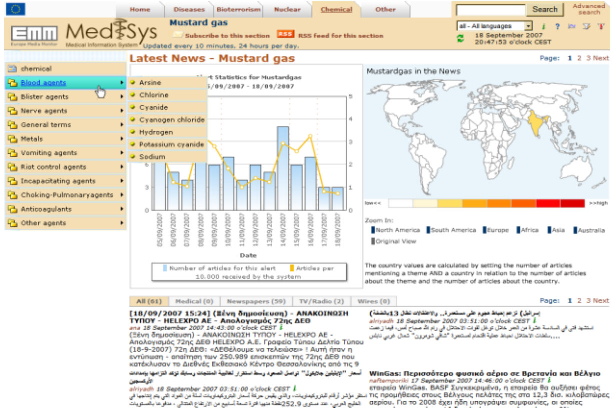
George W. Bush (45)
Ban Ki Moon (21)
Luciano Pavarotti (18)
Vladimir Putin (12)

This Week's New Stories

Felix makes landfall in Nicaragua
September 5, 2007 - September 6, 2007
Fossett remains missing in Nevada desert
September 4, 2007 - September 6, 2007
Luciano Pavarotti 1935-2007
September 5, 2007 - September 6, 2007
Millions face London Tube strike chaos
September 3, 2007 - September 5, 2007
UN accuses Rwanda of arming Congo rebel leader
September 1, 2007 - September 6, 2007
Apple overhauls entire iPod line
August 30, 2007 - September 6, 2007
Mattel 'set to recall more toys'
August 30, 2007 - September 6, 2007
Read more...

This Month's New Stories

Terror as fires sweep Greece
FIRES are still raging across Greece as 56 people have been left dead
August 25, 2007 - September



Mustard gas

Subscribe to this section RSS RSS feed for this section

Updated every 10 minutes, 24 hours per day.

18 September 2007
20:47:53 o'clock CEST

Latest News - Mustard gas

Page: 1 2 3 Next

chemical
Blood agents
Blister agents
Nerve agents
General terms
Metals
Vomiting agents
Riot control agents
Incapacitating agents
Choking-Pulmonaryagents
Anticoagulants
Other agents

Arsine
Chlorine
Cyanide
Cyanogen chloride
Hydrogen
Potassium cyanide
Sodium

Mustardgas in the News

The country values are calculated by setting the number of articles mentioning a theme AND a country in relation to the number of articles about the theme and the number of articles about the country.

All (61) | Medical (0) | Newspapers (59) | TV/Radio (2) | Wires (0)

Page: 1 2 3 Next

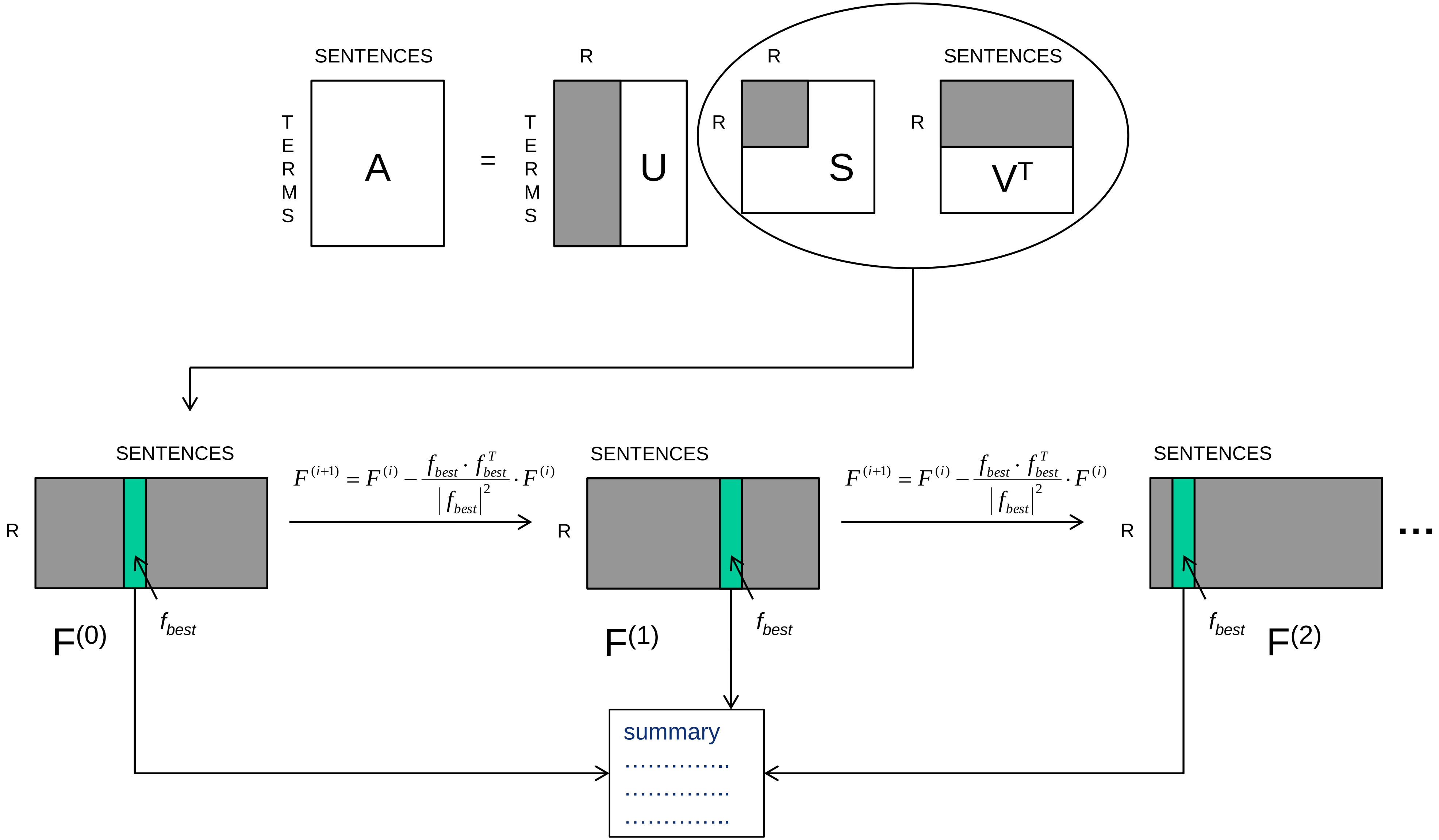
[18/09/2007 15:24] (Ξένη δημοσίευση) - ΑΝΑΚΟΙΝΩΣΗ ΤΥΠΟΥ - HELEXPO AE - Απολογισμός 72ης ΔΕΘ
ana 18 September 2007 14:43:00 o'clock CEST
(Ξένη δημοσίευση) - ΑΝΑΚΟΙΝΩΣΗ ΤΥΠΟΥ - HELEXPO AE - Απολογισμός 72ης ΔΕΘ HELEXPO Α.Ε. Γραφείο Τύπου Δελτίο Τύπου (18-9-2007) 72η ΔΕΘ: «ΔΕΘέλουμε να τελειώσει» ! Αυτή ήταν η εντύπωση - απαίτηση των 250.989 επισκεπτών της 72ης ΔΕΘ που κατέκλυσαν το Διεθνές Εκθεσιακό Κέντρο Θεσσαλονίκης από τις 9

[illegible]
alriyadh 18 September 2007 03:51:00 o'clock CEST
[illegible]

[illegible]
alriyadh 18 September 2007 03:51:00 o'clock CEST
[illegible]

WinGas: Περισσότερο φυσικό αέριο σε Βρετανία και Βέλγιο
naftemporiki 17 September 2007 14:46:00 o'clock CEST
εταιρεία WinGas. BASF Συγκεκριμένα, η εταιρεία θα αυξήσει φέτος τις προμήθειες στους Βέλγους πελάτες της στα 12,3 δισ. κιλοβατώρες αερίου. Για το 2008 έχει ήδη υπογράψει συμφωνίες, οι οποίες

Extractive Summarizer based on LSA

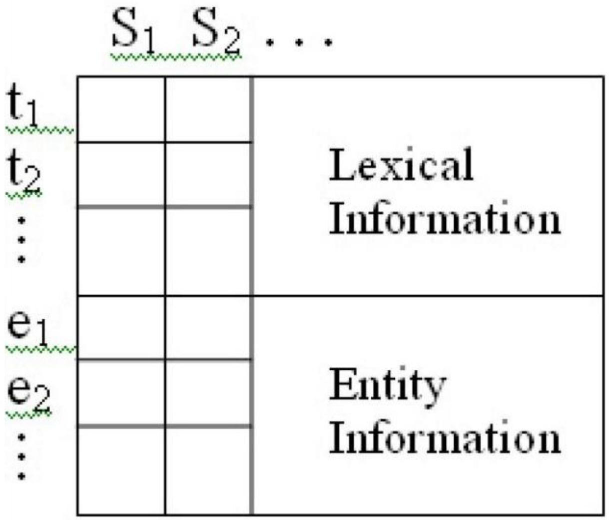


$$F^{(i+1)} = F^{(i)} - \frac{f_{best} \cdot f_{best}^{T}}{|f_{best}|^{2}} \cdot F^{(i)}$$

$$F^{(i+1)} = F^{(i)} - \frac{f_{best} \cdot f_{best}^{T}}{|f_{best}|^{2}} \cdot F^{(i)}$$

Extended knowledge representation

- Entity (Person/Organization/Location) names are not treated only lexically
- Entities get more weight
- Sentences sharing the same entities are closer to each other in the LSA space



Capturing aspects

- We used our **event extraction system** (NEXUS) + **a tool for learning of semantic classes** (Ontopopulis)
- The extracted information is **combined with co-occurrence information from LSA**
- Event extraction system (NEXUS)

  "All the 20 people taken hostage by armed pirates were safe."

  Extracted slots:

  event type (`kidnapping`), *victims* (`20 people`), *perpetrator* (`pirates`)

  - Captured TAC aspects: what happened, who affected, perpetrators
- Automatically learnt Lexica (Ontopopulis)
  - Sample from lexicon for countermeasures:

    operation, rescue operation, rescue, evacuation, treatment, assistance, relief, military operation, police operation, security operation, aid
  - Captured TAC aspects: damages, countermeasures, charges, what (resources)

Temporal analysis

- **Types of temporal expressions we cover:**
  - numerical vs. non-numerical: *03/18/2010 vs. on the fifth of December 2009*
  - fully specified vs. underspecified: *on the fifth of December 2009 vs.* in March 2002
  - absolute vs. relative vs. deictic: in March 2002 vs. *in March* vs. *last month*
  - simple vs. compound: *a year before last Monday*
  - discrete vs. fuzzy: *three days ago* vs. *in a few months*
- **Recognition and normalization**
  - Relative expression: anchor selection starts with the article date and is updated
- **3 applications**
  - Capturing the WHEN aspect
    - the most frequent normalized time
  - Identification of update sentences
    - Larger weight for sentences in which at least one of the temporal intervals is in an "after", "overlapped by" or "finishes" relation with the reference one (the date of the most recent article of the initial set).
  - Sentence ordering
    - The best sentence comes first, following sentences ordered by date/time (found in the particular sentence or in the preceding context or article date)

MT-based sentence compression and reconstruction

- **Motivation**
  - To generate summaries from our summary representation without recurring to simple sentence extraction.
  - Human summaries contain more and shorter sentences than system summaries (in TAC'09 – 6 vs. 4)

- **Approach**
  - Select the most important sentences
  - Leave only the important terms in the summary sentences
    - LSA + language-model
    - Parameterized compression rate
    - Output: sequence of important words
  - Reconstruct the sentences using the noisy-channel model
    - Monolingual phrase based statistical model in machine translation
    - "Translate" = to find the most probable target sentence by inserting new words and reproducing the inflected surface forms of the source words

Steinberger Josef, Marco Turchi, Mijail Kabadjov, Nello Cristianini & Ralf Steinberger (2010). Wrapping up a Summary: from Representation to Generation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'2010), pp. 382-286. Uppsala, Sweden, 11-16 July.

An example of the generative approach

Original sentence:

A Palestinian suicide bomber detonated an explosive belt at a commercial center in Dimona on Monday morning, killing an Israeli woman and wounding at least eight others.

Compressed and reconstructed:

A **Palestinian suicide bomber detonated** an **explosive commercial center** in **Dimona** and on **Monday morning, killing** an **Israeli.**

TERM	a	palestinian	suicide	bomber	detonated	an	explosive	belt	at	a	commercial	center	in	dimona	on	monday	morning	,	killing	an	israeli	woman	and	wounding	at	least	eight	others	.
LSA score	0	.32	.66	.64	.26	0	.26	.21	0	0	.24	.23	0	.56	0	.30	.11	0	.24	0	1	.16	0	.15	0	0	0	0	0
1-gram	0	.59	.26	.07	.07	0	.15	.12	0	0	.66	.83	0	.01	0	1	.17	0	.37	0	.70	.49	0	.03	0	0	0	0	0
2-gram	0	.02	.37	.37	.01	0	.01	.01	0	0	.05	.05	0	0	0	1	1	0	0	0	.02	.02	0	0	0	0	0	0	0
3-gram	0	.02	.31	.31	1	1	1	0	0	0	0	.00	.00	.00	.00	.00	.02	.02	.03	.03	.03	.03	.03	.03	0	0	0	0	0
4-gram	0	.00	.00	.17	1	1	1	1	.00	.00	.00	.00	.00	.00	.00	.30	.30	.30	.30	.00	.00	.00	.00	.00	0	0	0	0	0
Combined	**0**	**.28**	**.43**	**.38**	**.26**	**.14**	**.28**	**.13**	**0**	**0**	**.27**	**.31**	**0**	**.22**	**0**	**.57**	**.29**	**.01**	**.20**	**.00**	**.57**	**.19**	**.00**	**.07**	**0**	**0**	**0**	**0**	**0**

Results in the guided task

Initial summaries

50 submissions in total

ID	Overall responsiveness	Linguistic quality	Pyramid score	Number of repetitions
25 (the best run in Overall resp.)	3.159 (1)	3.341 (6)	0.440 (10)	1.409 (17/25)
22 (the best run in Pyramid score)	3.136 (2)	3.432 (5)	0.477 (1)	1.045 (7/25)
37 (sentence extraction)	**2.977 (12)**	**3.455 (4)**	**0.412 (23)**	**0.864 (2/25)**
6 (+ compression/paraphrasing)	2.341 (43)	2.318 (42)	0.311 (42)	0.568 (-/25)
2 (baseline - MEAD)	2.841 (27)	2.818 (30)	0.362 (32)	1.432 (-/25)
1 (baseline - LEAD)	2.500 (37)	3.205 (7)	0.304 (45)	0.455 (-/25)

Top 25 systems

Update summaries

ID	Overall responsiveness	Linguistic quality	Pyramid score	Number of repetitions
35 (the best run in Overall resp.)	2.591 (1)	2.818 (24)	0.342 (4)	0.818 (19/25)
9 (the best run in Pyramid score)	2.523 (5)	2.659 (34)	0.353 (1)	0.409 (3/25)
37 (sentence extraction)	**2.205 (31)**	**3.250 (6)**	**0.291 (21)**	**0.25 (1/25)**
6 (+ compression/paraphrasing)	1.864 (45)	2.159 (44)	0.176 (44)	0.295 (-/25)
2 (baseline - MEAD)	2.114 (35)	2.841 (22)	0.284 (24)	0.568 (10/25)
1 (baseline - LEAD)	2.091 (37)	3.455 (1)	0.237 (36)	0.364 (-/25)

Using Parallel Corpora for Multilingual Summarization Evaluation

- Data – Project Syndicate (http://www.project-syndicate.org/)
  - Commentaries and analyses of important world event
  - Original text human-translated into various other languages (En, Fr, Es, De, Ru, Ar, Cz)
- Sentence-aligned (91.7% one-to-one alignments) parallel corpus
- Manual selection of the most important sentences
  - 4 annotators / 78% inter-annotator agreement (at least two annotators)
- **Projecting the sentence selection to various target languages**
- Available for download: http://langtech.jrc.it/JRC_Resources.html#Summarisation-evaluation
- Conclusions:
  - LSA-based summarizer selects different sentences in different languages (~40% agreement);
  - Its performance is comparable across languages
  - Introducing entities improves short summaries and leads to selection of more similar content among languages

Turchi Marco, Josef Steinberger, Mijail Kabadjov & Ralf Steinberger (2010). Using parallel corpora for multilingual (multi-document) Summarisation Evaluation. Conference on Multilingual and Multimodal Information Access Evaluation (CLEF'2010). Padua, Italy, 20-23 September 2010. Springer Lecture Notes for Computer Science LNCS.

JRC
EUROPEAN COMMISSION

Multilingual task – human grades

ipsc
Institute for the Protection and Security of the Citizen

ARABIC

ID	Score	Significance
B	~4.10	A
1	3.77	A B
9	3.73	A B
8	3.70	A B
3	**3.43**	**A B C**
7	3.30	B C
10	3.20	B C
2	3.10	B C
6	3.10	B C
4	2.77	C D
C	~2.20	D E
A	~1.90	E

CZECH

ID	Score	Significance
B	4.89	A
C	4.80	A
A	4.73	A
D	4.50	A
3	***3.40***	***B***
9	3.30	B
1	30	B C
2	2.70	C
10	2.68	C
7	2.20	D
4	1.48	E

ENGLISH

ID	Score	Significance
A	~4.40	A
B	~4.30	A B
3	**3.57**	**A B S**
C	~3.53	B S T
2	3.53	B S T V
1	3.20	S T VX
10	3.20	S T VX
5	2.73	S T VXY
8	2.47	T VXYZ
6	2.40	VXYZ
9	2.27	XYZ
7	2.10	YZ
4	1.80	Z

FRENCH

ID	Score	Significance
F	~4.70	A B C
C	~4.30	A
A	~4.25	A B
D	~4.20	A B
E	~3.90	A B C
B	~3.50	B C
3	**3.23**	**C**
1	2.30	D
2	2.20	D
6	2.20	D
10	2.10	D
7	2.07	D
9	2.03	D
5	1.90	D
4	1.33	E

GREEK

ID	Score	Significance
A	~4.40	A
B	~4.10	A B
3	**3.63**	**B C**
C	~3.50	B C
2	3.33	B C
10	3.30	B C
9	3.13	C
1	3.00	C
7	2.10	D
4	1.97	D

HEBREW

ID	Score	Significance
A	~4.60	A
3	**3.87**	**A B**
B	~3.60	B C
C	~3.60	B C
1	3.29	B C
2	3.29	B C
9	3.16	B C
7	3.06	C
10	3.03	C
4	2.19	D

HINDI

ID	Score	Significance
B	~4.70	A
A	~4.20	A B
C	~4.00	B
5	2.73	C
2	2.53	C
1	2.50	C
3	**2.47**	**C**
7	2.00	D
6	1.90	D E
10	1.83	D E
9	1.80	D E
4	1.40	E

Multilingual task – adjusting the grades

Raw grades

ID	Czech	English	French	Arabic	Greek	Hebrew	Hindi	Average
1	3.00	3.20	2.30	3.77	3.00	3.29	2.50	3.01
2	2.70	3.53	2.20	3.10	3.33	3.29	2.53	2.96
3	**3.40(1)**	**3.57(1)**	**3.23(1)**	**3.43(4)**	**3.63(1)**	**3.87(1)**	**2.47(4)**	**3.37(1)**
4	1.48	1.80	1.33	2.77	1.97	2.19	1.40	1.85
5		2.73	1.90				2.73	2.46
6		2.40	2.20	3.10			1.90	2.40
7	2.20	2.10	2.07	3.30	2.10	3.06	2.00	2.40
8		2.47		3.70				3.08
9	3.30	2.27	2.03	3.73	3.13	3.16	1.80	2.78
10	2.68	3.20	2.10	3.20	3.30	3.03	1.83	2.76

Length-aware human grades

ID	Czech	English	French	Arabic	Greek	Hebrew	Hindi	Average
1	3.00	3.10	2.28	3.77	3.00	3.29	2.50	2.99
2	2.70	3.53	2.20	3.10	3.33	3.29	2.53	2.96
3	**3.15(2)**	**3.31(2)**	**2.93(1)**	**3.10(7)**	**3.25(3)**	**3.56(1)**	**2.20(4)**	**3.07(1)**
4	1.48	1.80	1.33	2.76	1.97	2.19	1.27	1.83
5		2.63	1.69				2.60	2.31
6		1.98	1.62	2.76			1.64	2.00
7	2.19	2.09	2.05	3.30	2.10	3.06	1.99	2.40
8		2.47		3.66				3.06
9	3.30	2.27	2.03	3.73	3.13	3.16	1.80	2.78
10	2.68	3.20	2.10	3.20	3.30	2.85	1.75	2.73

Multilingual task – ROUGE-2 and AutoSummENG

ROUGE-2

ID	Czech	English	French	Arabic	Greek	Hebrew	Hindi	**Average**
1	0.173	0.121	0.153	0.120	0.061	0.051	0	0.097
2	0.190	0.170	0.197	0.138	0.149	0.095	0.008	0.135
3	***0.199(1)***	**0.173(1)**	**0.201(1)**	**0.158(1)**	**0.100(3)**	**0.129(1)**	**0.058(1)**	**0.145(1)**
4	0.178	0.151	0.159	0.139	0.128	0.082	0.025	0.123
5		0.136	0.118				0.033	
6		0.106	0.100	0.137			0.008	
7	0.138	0.096	0.150	0.106	0.049	0.090	0	0.090
8		0.121		0.147				
9	0.139	0.109	0.129	0.125	0.072	0.102	0	0.097
10	0.351	0.251	0.285	0.233	0.145	0.222	0	0.212

AutoSummENG

ID	Czech	English	French	Arabic	Greek	Hebrew	Hindi	**Average**
1	0.361	0.344	0.387	0.297	0.306	0.277	0.240	0.316
2	0.373	0.386	0.414	0.369	0.375	0.327	0.286	0.361
3	***0.430(1)***	**0.426(1)**	**0.466(1)**	**0.483(1)**	**0.372(2)**	**0.368(1)**	**0.275(2)**	**0.403(1)**
4	0.370	0.379	0.367	0.383	0.330	0.313	0.199	0.334
5		0.350	0.358				0.249	
6		0.349	0.353	0.340			0.220	
7	0.316	0.311	0.382	0.261	0.287	0.280	0.199	0.291
8		0.332		0.305				
9	0.312	0.304	0.336	0.282	0.291	0.272	0.207	0.286
10	0.689	0.548	0.595	0.666	0.524	0.537	0.361	0.560

Conclusions

- **On the road towards multilingual summarization**
  - Limiting dependence on a particular language
  - The core LSA approach uses only list of stopwords
  - Entity recognition – 20 languages
  - Event extraction – 8 languages
  - Temporal analysis – 4 languages
- **Guided task**
  - Temporal analysis added
  - Low redundancy + good linguistic quality, average content
  - Compression/paraphrasing – language-independent but we are not there yet
- **Multilingual task**
  - Comparable to some humans for 5 languages
  - Well performing compared to other systems