Multilingual Multi-document Summarization Task in MultiLing 2015

George Giannakopoulos[1,2]

[1]NCSR Demokritos, Greece
ggianna@iit.demokritos.gr
[2]SciFY NPC, Greece

September 2015

Task description

An event sequence:

- set of atomic (self-sufficient) event descriptions
- sequenced in time
- share actors, location of occurence or other important factor

- Generate a single, fluent, representative summary
- ...from a set of documents describing an *event sequence*
- ...language for document set within a given range
- ...output summary should be (240-)250 words

Dataset

- Human created
- Multi-lingual
- News
- Freely available
- Containing event sequences
- Plain text

Solution

- WikiNews (`http://www.wikinews.org`)
- Translation
- Preprocessing

Corpus update

- Clean-up of previous corpus (10-15 topics / language)
- Limited human evaluation (English and Greek complete so far)

Model summaries

- 3 summarizers per topic and language
- Keep human subjectivity related to important aspects
- Use the minimum possible guidelines
  - Self-sufficient, clearly written text
  - ...providing no external information
  - ...fluent, easily readable language

Model summaries

- 3 summarizers per topic and language
- Keep human subjectivity related to important aspects
- Use the minimum possible guidelines
  - Self-sufficient, clearly written text
  - ...providing no external information
  - ...fluent, easily readable language
- Clean-up identical human summaries

Manual Evaluation

- Partial overlap of evaluators with summarizers.
- *No self-assessment*.
- At least three evaluations per summary.

Guidelines

- Read source documents at least once
- Give a grade between 1 and 5 (Overall Responsiveness: OR)
- Content and fluency equally important

Manual Evaluation

- Partial overlap of evaluators with summarizers.
- *No self-assessment*.
- At least three evaluations per summary.
- Reduced to 2 due to overlap in few cases.

Guidelines

- Read source documents at least once
- Give a grade between 1 and 5 (Overall Responsiveness: OR)
- Content and fluency equally important

Methods used

- ROUGE
- N-gram graph Based: MeMoG, AutoSummENG, NPower

Finding

Pre-processing and encoding can cause significant evaluation changes

Participants

SysID	Participant and System name
MMS1	UJF-Grenoble
MMS2	UWB
MMS3	ExB
MMS5	ESI-AllSummarizer
MMS8	IDA-OCCAMS
MMS9	GiauUngVan
MMS11	SCE-Poly
MMS12	BUPT-CIST
MMS13	BGU-MUSE
MMS15	NCSR/SCIFY-NewSumRerank

Table : List of participants

Arabic

Groups, Treatments and means		
a	MMS2	0.2235
a	MMS11a_Litvak_Vanetik_system1_MMS_poly_G	0.2167
a	MMS8a_FASTTF_OCCAMS_Priority1	0.202
ab	MMS8c_FAST_rank24_OCCAMS_Priority3	0.1875
ab	MMS13a-MUSE	0.1803
ab	MMS13b-MUSE	0.1803
ab	MMS15	0.1801
ab	MMS8b_BasisTF_OCCAMS_Priority2	0.1798
ab	MMS5a	0.172
ab	MMS11b_Litvak_Vanetik_system2_MMS_poly_Kr	0.1704
ab	MMS3	0.1667
ab	MMS5b	0.1666
b	MMS12-CIST	0.106

Table : Tukey HSD on Arabic — Baseline ≈ 0.13

Chinese

```
Groups, Treatments and means
a       MMS3                                  0.1314
ab      MMS5a                                 0.1038
abc     MMS2                                  0.1019
abcd    MMS5b                                 0.08431
 bcd    MMS8a_FASTTF_OCCAMS_Priority1         0.04379
  cd    MMS12-CIST                            0.03614
   d    MMS8c_FAST_rank24_OCCAMS_Priority3    0.02558
   d    MMS8b_BasisTF_OCCAMS_Priority2        0.02093
   d    MMS15                                 0.01987
```

Table : Tukey HSD on Chinese — Baseline ≈ 0.02

Czech

```
Groups, Treatments and means
a     MMS2                                  0.2312
ab    MMS8a_FASTTF_OCCAMS_Priority1         0.2265
ab    MMS8b_BasisTF_OCCAMS_Priority2        0.2186
ab    MMS8c_FAST_rank24_OCCAMS_Priority3    0.2089
ab    MMS3                                  0.1971
ab    MMS5a                                 0.1931
abc   MMS12-CIST                            0.1781
 bc   MMS5b                                 0.1673
  c   MMS15                                 0.1292
```

Table : Tukey HSD on Czech — Baseline ≈ 0.15

English

Groups, Treatments and means		
a	MMS2	0.222
a	MMS8a_FASTTF_OCCAMS_Priority1	0.2185
a	MMS8c_FAST_rank24_OCCAMS_Priority3	0.2159
a	MMS8b_BasisTF_OCCAMS_Priority2	0.2129
a	MMS15	0.2004
a	MMS1a	0.1988
a	MMS1b	0.196
a	MMS5b	0.1944
a	MMS5a	0.1936
a	MMS13b-MUSE	0.1849
a	MMS3	0.1848
a	MMS1c	0.1817
a	MMS13a-MUSE	0.1801
a	MMS11a_Litvak_Vanetik_system1_MMS_poly_G	0.18
a	MMS11b_Litvak_Vanetik_system2_MMS_poly_Kr	0.1794
a	MMS12-CIST	0.1651

Table : Tukey HSD on English — Baseline ≈ 0.18

French

Groups, Treatments and means		
a	MMS2	0.2661
ab	MMS8b_BasisTF_OCCAMS_Priority2	0.243
ab	MMS8a_FASTTF_OCCAMS_Priority1	0.2257
ab	MMS8c_FAST_rank24_OCCAMS_Priority3	0.2241
ab	MMS1a	0.2108
ab	MMS1b	0.2018
ab	MMS3	0.1982
ab	MMS5a	0.1911
ab	MMS1c	0.1863
ab	MMS5b	0.1851
b	MMS15	0.1743
b	MMS12-CIST	0.1639

Table : Tukey HSD on French — Baseline ≈ 0.20

Greek

```
Groups, Treatments and means
a     MMS8a_FASTTF_OCCAMS_Priority1          0.1823
ab    MMS8c_FAST_rank24_OCCAMS_Priority3     0.1809
ab    MMS2                                   0.1727
ab    MMS1a                                  0.1713
ab    MMS8b_BasisTF_OCCAMS_Priority2         0.17
ab    MMS15                                  0.1698
ab    MMS3                                   0.1625
ab    MMS12-CIST                             0.1513
ab    MMS1c                                  0.1492
ab    MMS5b                                  0.1368
ab    MMS1b                                  0.1337
 b    MMS5a                                  0.1292
```

Table : Tukey HSD on Greek — Baseline ≈ 0.14

Hebrew

```
Groups, Treatments and means
a    MMS15                                      0.1863
a    MMS8b_BasisTF_OCCAMS_Priority2             0.1859
a    MMS2                                       0.1832
ab   MMS5a                                      0.1721
ab   MMS13a-MUSE                                0.1693
ab   MMS13b-MUSE                                0.1693
ab   MMS8a_FASTTF_OCCAMS_Priority1              0.1683
ab   MMS11a_Litvak_Vanetik_system1_MMS_poly_G   0.1624
ab   MMS5b                                      0.1596
ab   MMS8c_FAST_rank24_OCCAMS_Priority3         0.159
ab   MMS3                                       0.1441
ab   MMS11b_Litvak_Vanetik_system2_MMS_poly_Kr  0.1324
 b   MMS12-CIST                                 0.1214
```

Table : Tukey HSD on Hebrew — Baseline ≈ 0.14

Hindi

```
Groups, Treatments and means
a     MMS8b_BasisTF_OCCAMS_Priority2          0.1573
a     MMS8a_FASTTF_OCCAMS_Priority1           0.1491
a     MMS2                                    0.1479
a     MMS8c_FAST_rank24_OCCAMS_Priority3      0.1461
a     MMS3                                    0.1377
a     MMS12-CIST                              0.1329
 b    MMS5a                                   0.06968
 b    MMS5b                                   0.05332
 b    MMS15                                   0.03744
```

Table : Tukey HSD on Hindi — Baseline ≈ 0.12

Romanian

```
Groups, Treatments and means
a    MMS8a_FASTTF_OCCAMS_Priority1          0.2165
a    MMS2                                   0.2145
a    MMS8c_FAST_rank24_OCCAMS_Priority3     0.2055
a    MMS8b_BasisTF_OCCAMS_Priority2         0.2051
a    MMS3                                   0.1779
a    MMS5a                                  0.1765
a    MMS15                                  0.1651
a    MMS5b                                  0.1622
a    MMS12-CIST                             0.1557
```

Table : Tukey HSD on Romanian — Baseline ≈ 0.17

Spanish

a	MMS2	0.2516
a	MMS8b_BasisTF_OCCAMS_Priority2	0.2454
a	MMS8a_FASTTF_OCCAMS_Priority1	0.2297
a	MMS8c_FAST_rank24_OCCAMS_Priority3	0.2272
a	MMS3	0.2166
a	MMS5a	0.2051
a	MMS15	0.2038
ab	MMS5b	0.1997
b	MMS12-CIST	0.1373

Table : Tukey HSD on Spanish — Baseline ≈ 0.19

(An) Overall ranking

SysID	Total Rank
MMS2	102
MMS8a_FASTTF_OCCAMS_Priority1	92
MMS8b_BasisTF_OCCAMS_Priority2	82
MMS8c_FAST_rank24_OCCAMS_Priority3	76
MMS3	53
MMS5a	52
MMS15	50
MMS5b	36
MMS1a	28
MMS13b-MUSE	24
MMS12-CIST	22
MMS11a_Litvak_Vanetik_system1_MMS_poly_G	21
MMS13a-MUSE	21
MMS1b	19
MMS1c	13
MMS11b_Litvak_Vanetik_system2_MMS_poly_Kr	8

Table : Sum of ranks over all languages — Penalizes participation in few languages

Method

- Online tool (Thank you SciFY people and George Kiomourtzis!)
- Evaluation based on human summary (and not source texts)
- Per language sampling of document pairs
- One evaluation per pair
- Statistics meaningful on the language level

Evaluation ongoing

We need your help to complete this! Please contact me, if you want to help in your language.
Check this website.

Current state and challenges

- Doing better than baseline (based on automatic evaluation)
- Language-specific aspects in automatic evaluation can pose problems
- Need for robust language-independent automatic evaluation
- Need more community effort for the human evaluation on several languages

Proposals for the future

- Re-activate evaluation task
- Large-scale crowd-sourced evaluation
- Online services for standardized evaluation on the corpus

More in the planning session...

Thank you!