Overview of the MultiLing Pilot in TAC 2011

George Giannakopoulos[1]

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

November 2011

Outline

Multilinguality

- News
- Blogs
- Search results
- Automatic translation

Brief history of DUC/TAC domains

- Single document summarization
- Multi-document summarization (Update, Guided, Opinion, ...)
- Cross-lingual summarization

Something appears to be missing...

The missing piece: MultiLing

Create summaries regardless of underlying language on document sets that use the same (possibly unknown) language.

MultiLing aim

- Detect multi-document summarization (MMS) research
- Learn about MMS algorithms
- Learn about multilingual reusable resources
- Quantify performance
- Check existing automatic measures

Outline

Task definition

- 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

An event Sequence

...is a set of atomic (self-sufficient) event descriptions, sequenced in time, that share main actors, location of occurence or some other important factor. Event sequences may refer to topics such as a natural disaster, a crime investigation, a set of negotiations focused on a single political issue, a sports event.

Dataset

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

Solution

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

Mini-pilot for effort estimation

- Small scale corpus (2 topics)
- Everything was timed
- Questions would be noted

Lesson

Always do a mini-pilot, note everything, do follow-up meetings.

Overview of full corpus creation

- Determine topics (10 topics / language)
- Translate documents (10 docs / topic)
- Produce model summaries (3 models / topic)

Determine topics

- Use metadata (WikiNews categories)
- Verify existence of event sequence
- Cover several different news types (e.g., politics, environment, sports)
- Find at least 10 documents per topic

Translate documents

- Sentence alignment
- Keep original meaning
- Produce readable, fluent text
- Translation verified

Lesson

Difficult, error-prone, subjective, high cost process.

Summarizing

- 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

Lesson

Few guidelines are better than a lot.

Types of evaluation

- Automatic (ROUGE, AutoSummENG)
- Manual (Overall Responsiveness)

Automatic Methods

- ROUGE (ROUGE-1, 2, SU-4), word n-gram matching, allows gaps
- AutoSummENG — Merged Model Graph (MeMoG), character n-gram co-occurence, merged representation

Not (too) strongly correlated. Possibly describing slightly different aspects.

Manual Evaluation Guidelines

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

Guidelines continued

We consider a text to be worth a 5, if it appears to cover all the important aspects of the corresponding document set using fluent, readable language. A text should be assigned a 1, if it is either unreadable, nonsensical, or contains only trivial information from the document set.

Outline

Overview

- Original aim: 3 groups per language

Overview

- Original aim: 3 groups per language
- Achieved: 8+1 groups

Overview

- Original aim: 3 groups per language
- Achieved: 8+1 groups
- Original aim: 5 languages

Overview

- Original aim: 3 groups per language
- Achieved: 8+1 groups
- Original aim: 5 languages
- Achieved: 7 languages

Baseline — Topline

global baseline system (ID9) , vector space, bag-of-words, highest cosine similarity to the centroid of documents.

global topline system (ID10) uses the model summaries, produces random summaries by combining sentences, find the one closest to the Merged Model Graph of the models.

Our champions

Participant	System ID	Arabic	Czech	English	French	Greek	Hebrew	Hindi	Notes
CIST	ID1	✓	✓	✓	✓	✓	✓	✓	Peer
CLASSY	ID2	✓	✓	✓	✓	✓	✓	✓	Peer
JRC	ID3	✓	✓	✓	✓	✓	✓	✓	Coorg (Czech)
LIF	ID4	✓	✓	✓	✓	✓	✓	✓	Coorg (French)
SIEL_IIITH	ID5			✓	✓			✓	Coorg (Hindi)
TALN_UPF	ID6	✓		✓	✓			✓	Peer
UBSummarizer	ID7	✓	✓	✓	✓	✓	✓	✓	Peer
UoEssex	ID8	✓		✓					Coorg (Arabic)
Baseline	ID9	Centroid baseline for all languages							Coorg (All)
Topline	ID10	Using model summaries for all languages							Coorg (All)

Lesson

The community *will* respond if you take the first step.

Evaluation aims

- Allow, but penalize, out-of-limit text sizes
- Measure per language performance
- Reward multi-lingual systems

Length-Aware Grading (LAG)

Given a summary S of length $|S|$ (in words) assigned a grade g, a lower word limit count l_{min} and an upper word limit count l_{max}:

$$LAG(g, S) = g * \left(1 - \frac{\max(\max(l_{min} - |S|, |S| - l_{max}), 0)}{l_{min}}\right)$$

Example

An excellent summary (graded with OR 5) with 120 words, would be assigned a LAG-OR grade of 2.5 (less than mediocre).

Combined Multi-lingual Performance (CMP)

$g_s(l)$ is the LAG grade of system s in a given language l from the full set of languages L:

$$\text{CMP}_s = \frac{\sum_{l \in L} g_s(l)}{|L|}$$

Non-participation implies a LAG value of 1.

Instability

System s participated in the set L_s of languages, $L_s \subset L$, and the st.dev. of its LAG grades in these languages is σ_s, then:

$$\text{Instability}_s = \frac{\sigma_s}{\sqrt{|L_s|}}$$

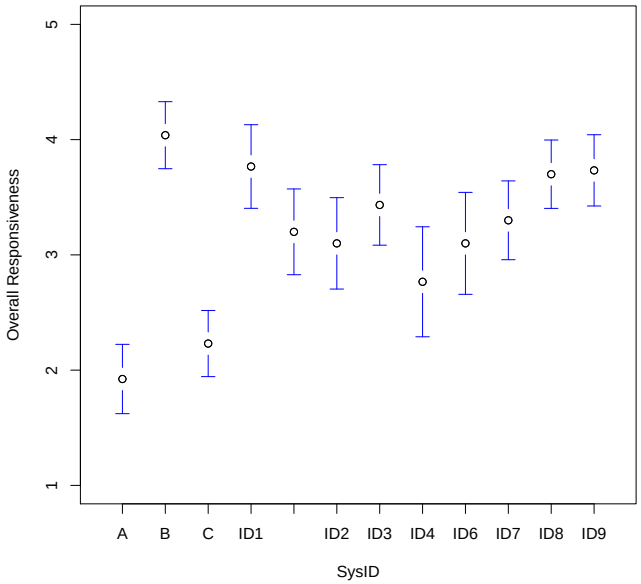
Higher instability indicates more uncertainty on future performance

Overview

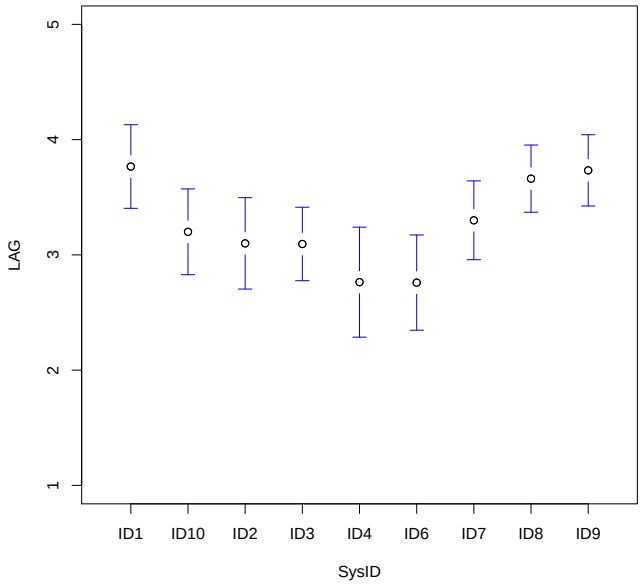
System	CMP	Instability
ID1 (CIST)	2.99	0.19
ID2 (CLASSY)	2.95	0.18
ID3 (JRC)	3.13	0.18
ID4 (LIF)	1.86	0.21
ID5 (SIEL_IIITH)	1.6	0.48
ID6 (TALN_UPF)	1.6	0.34
ID7 (UBSummarizer)	2.41	0.19
ID8 (UoEssex)	1.63	0.78
ID9 (Baseline)	2.81	0.27
ID10 (Topline)	2.71	0.22

Table: Combined Multi-lingual Performance and Instability per System

Per Language Overview — Arabic



Overall Responsiveness

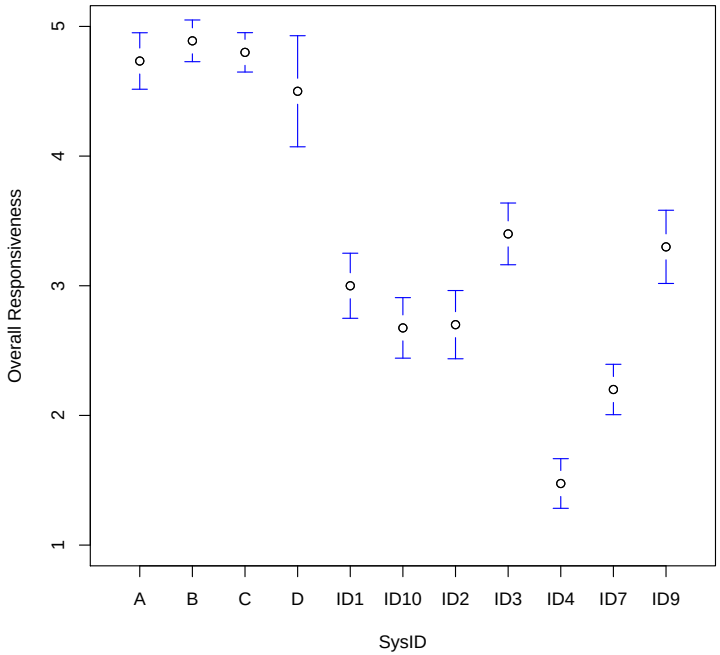


LAG (Systems only)

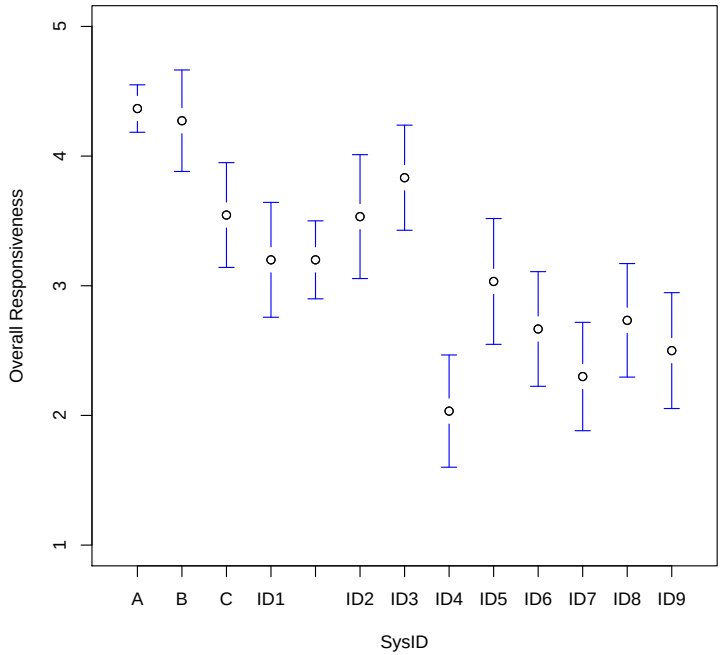
Lesson

Model summaries may be bad summaries. How does this influence evaluation?

Overall Responsiveness — Czech, English

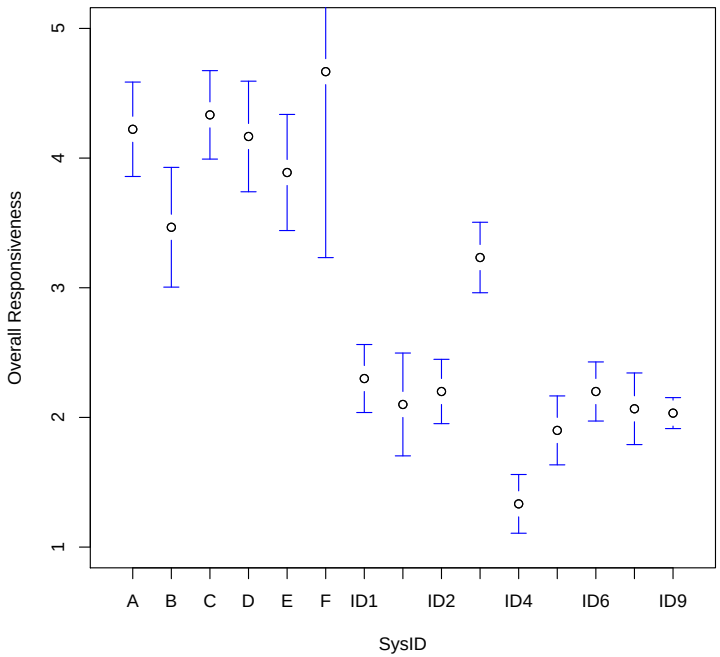


Czech

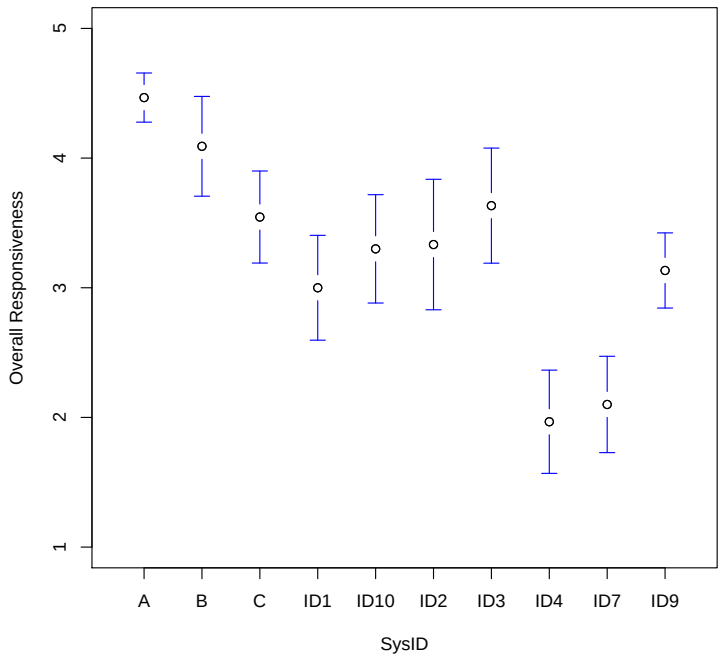


English

Overall Responsiveness — French, Greek

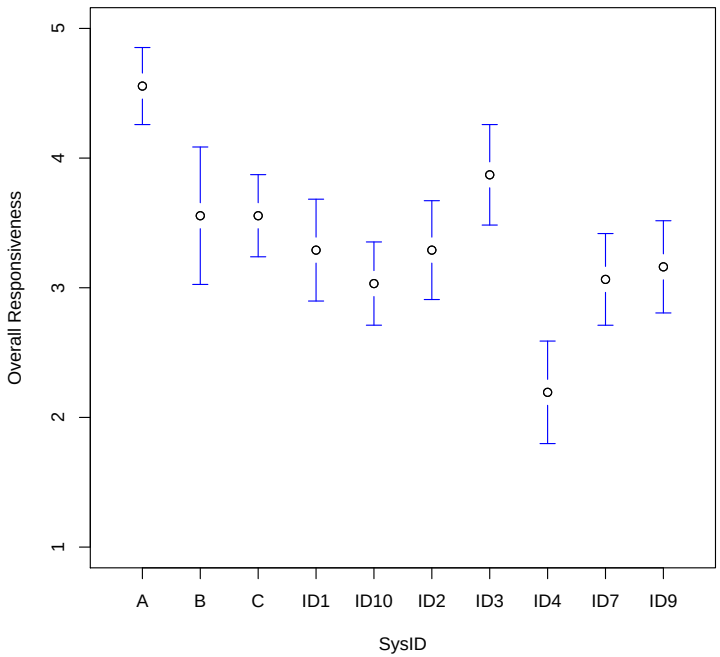


French

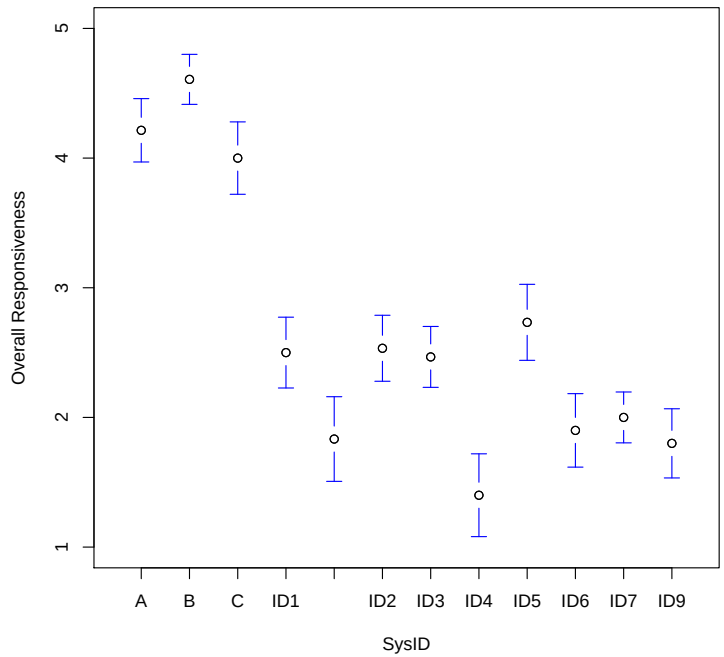


Greek

Overall Responsiveness — Hebrew, Hindi



Hebrew



Hindi

Summary of system performances

- Systems good enough for many languages
- Big variance across languages
- Human grades not always stable
- Human grades not always high

Correlations

Language	ROUGE2 to OR	MeMoG to OR	ROUGE2 to MeMoG
Arabic	0.25	-0.36	0.11
Czech	0.33	-0.04	0.24
English	**0.56**	**0.47**	**0.47**
French	0.42	0.37	**0.50**
Greek	0.14	0.33	0.24
Hebrew	0.52	0.05	-0.24
Hindi	0.18	0.33	0.13
All languages	0.12	0.12	**0.42**

Table: Correlation (Kendall's Tau) Between Gradings. Note: statistically significant results in **bold**.

Lesson

Much space for improvement. Negative examples can be good examples...

Outline

Community

- MMS Researchers are present
- MMS Researchers are active and collaborating

Community

- MMS Researchers are present
- MMS Researchers are active and collaborating
- Researchers need data and evaluation

Dataset

- Useful
- Publicly available
- A basis for future work
- Measured effort

From pilot to track

- Dataset
- Evaluation
- Support

Dataset

- Change of scale
  - More languages
  - More texts
- Dataset creation support software
- (Funded) Community work

Evaluation

- Larger dataset
- Use negative examples of summaries
- Optimize existing metrics
- Devise better metrics

Support

- TAC support
- Community support
- AIJ funding

Thank you!

Last lesson

United we stand, divided we fall... (attributed to Aesop, Greek Fabulist)
We stand. (TAC MultiLing Pilot Community)

Co-organizers:

- Ilias Zavitsanos, (NCSR Demokritos, Greece)
- Vasudeva Varma (IIT Hyderabad, India)
- Josef Steinberger (JRC, Italy in collaboration with the Univ. of West Bohemia, Czech Republic)
- Benoît Favre (LIF, France)
- Marina Litvak (Sami Shamoon College of Engineering, Israel)
- Mahmoud El - Haj (Univ. of Essex, UK)
- William Darling (Univ. of Guelph, Canada)