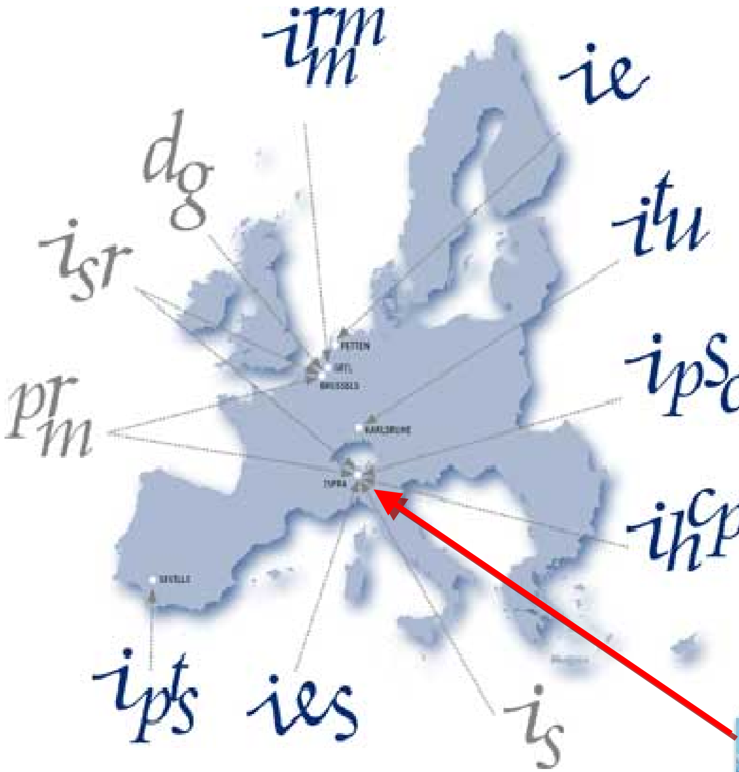
Using Parallel corpora for Multilingual (Multi-Document) summarisation Evaluation

Marco Turchi, Josef Steinberger, Mijail Kabadjov
and Ralf Steinberger
European Commission, Joint Research Centre
Institute for the Protection and Security of the Citizen,
Optima Action
{Name.Surname}@jrc.ec.europa.eu

Joint Research Centre - Who we are



BRUSSELS (BE)
- The Directorate General (**DG**)
- The Institutional and Scientific Relations Directorate (**ISR**)
- The Programme and Resource Management Directorate (**PRM**)

GEEL (BE)
- The Institute for Reference Materials and Measurements (**IRMM**)

KARLSRUHE (DE)
- The Institute for Transuranium Elements (**ITU**)

ISPRA (IT) Download the Ispra site Brochure (English - Italian)
- The Institute for the Protection and Security of the Citizen (**IPSC**)
- The Institute for Environment and Sustainability (**IES**)
- The Institute for Health and Consumer Protection (**IHCP**)
- The Ispra site Directorate (**IS**)

PETTEN (NL)
- The Institute for Energy (**IE**)

SEVILLE (E)
- The Institute for Prospective Technological Studies (**IPTS**)

"The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national."

The EMM Family of News Analysis Applications

- **NewsBrief**: current state of affairs, *breaking news* detection in real time
- **MedISys**: focusing on *health-related* news
- **NewsExplorer**: long-term, *cross-lingual* news analysis and *people and organization* monitor
- **EMM-Labs**: various data *visualization* and advanced *text processing* tools

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

Outline

- ***Motivation***
- Multilingual parallel evaluation data for summarisation
  - Corpus preparation
  - Human annotation/sentence selection of English documents
  - Automatic projection to all other languages
- Methodology for automatic summary evaluation
  - Comparison across languages
- Released Data
- Conclusion

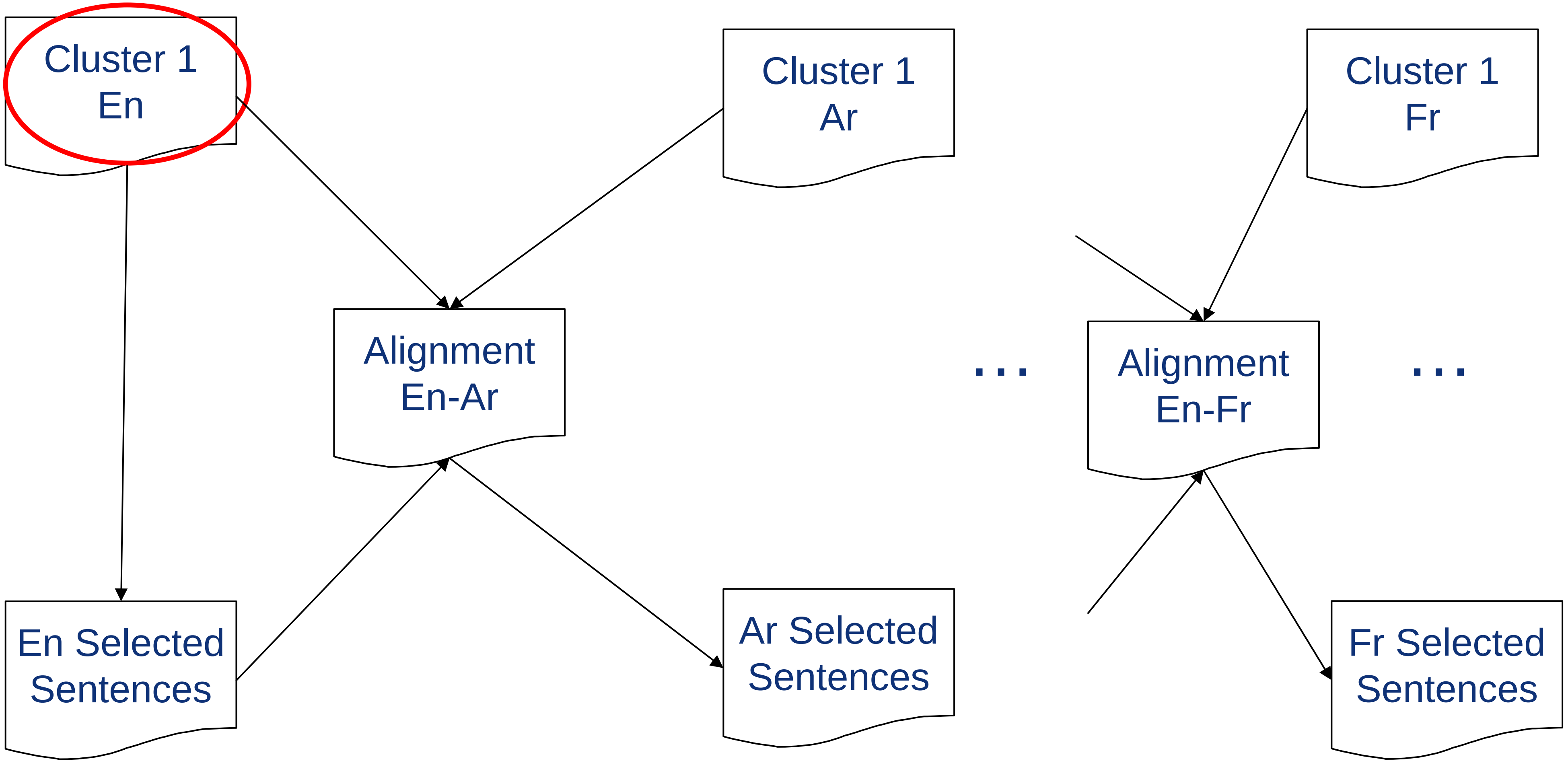
Using Parallel corpora for Multilingual (Multi-Document) summarisation Evaluation

- Given a collection of related documents, the goal of ***Automatic Summarisation*** is to ***produce a reliable and informative summary***.
- To evaluate the performance of each system, ***summaries*** need to be ***compared against a gold standard*** generally created by human beings. The most used automatic score is ROUGE.
- But:
  - generation process requires ***human interaction*** to ***extrapolate*** a short and coherent abstract;
  - this process is ***highly subjective***, ***time-consuming*** and ***expensive***;
  - human-annotated corpora are available for summarisation evaluation in ***English*** e.g. TAC;
  - even when such evaluation data exists for various languages, evaluation results are ***unlikely to be comparable across languages***.

- Focus on:
  - ***testing*** multi-document summarisation algorithms in ***languages other than English***;
  - ***comparing*** the results ***across languages;***
  - ***making*** the ***data available*** for research purposes.

- Main idea:
  - given a set of parallel documents in seven languages referring to a particular topic:
    - ***manually select*** the most representative sentences in one of the languages;
    - ***project to all other languages*** the selected sentences using the parallelism property of the documents.

Using Parallel corpora for Multilingual (Multi-Document) summarisation Evaluation



Using Parallel corpora for Multilingual (Multi-Document) summarisation Evaluation

- We propose:
  - a ***semi-automatic approach to generate corpora*** for research on multilingual summarisation taking advantage of the parallelism among documents in different languages;
  - an ***evaluation score*** based on different degrees of inter-annotator agreement between human annotators;
  - ***comparison*** of the performance of automatic summarisers ***on seven different languages***.
- The produced data are available for download.

Outline

- Motivation
- ***Multilingual parallel evaluation data for summarisation***
  - Corpus preparation
  - Human annotation/sentence selection of English documents
  - Automatic projection to all other languages
- Methodology for automatic summary evaluation
  - Comparison across languages
- Released Data
- Conclusion

Parallel Document Extraction

- A group of annotators with a Computer Science and a Linguistics background was chosen.

- Each human annotator was asked to select:
  - a topic from the Project Syndicate web page;
  - for each topic, a homogeneous set of five related English language documents (Only documents existing in at least English, French, Spanish, German, Arabic, Russian and Czech could be chosen);

- Remark:
  - http://www.project-syndicate.org/. Project Syndicate produces high quality commentaries of important world events. Each contributor produces a commentary in one language that is then human-translated into various languages.

- Annotators collected four topics:
  - ***Israeli-Palestinian conflict***, ***Malaria***, ***Genetics*** and ***Science-and-Society***.

Parallel Document Extraction

- For each topic, each document was downloaded and split into sentences:
  - average number of sentences per document was over 50.

- Every non-English sentence was aligned with the English version of the same sentence using Vanilla software.

- In total, there were:
  - 91.7% of one-to-one sentence alignments,
  - 3.4% two-to-one,
  - 4.49% one-to-two,
  - 0.2% two-to-two,
  - 0.3% zero-to-one.

Human Annotation

- All four annotators were asked to read and label, independently, all sentences from each English document of each cluster.

- After a pilot study, the definition of “summary-worthy” sentence was refined.

- Select sentences:
  - according to the cluster topic and document title;
  - that convey sufficient information;
  - that contain essential background and author’s point of view.

Inter-annotator agreement

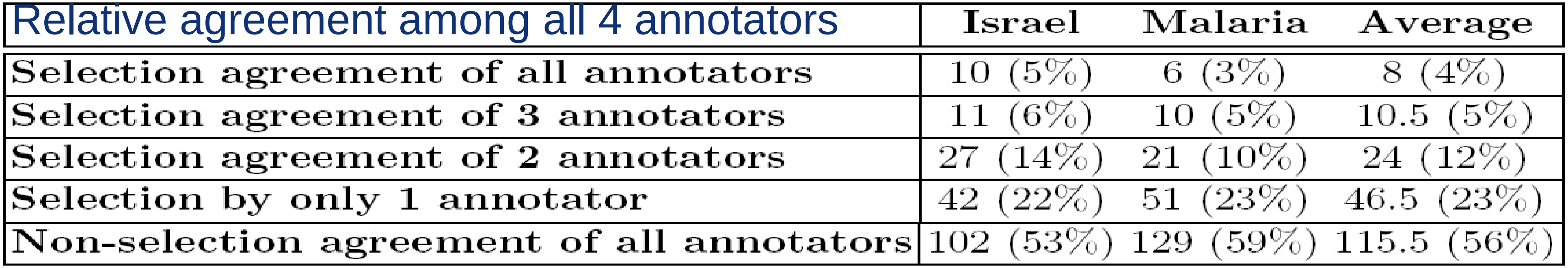
- Typically, two annotators do not produce the same gold standard annotation.

- Summary production is a very subjective task.

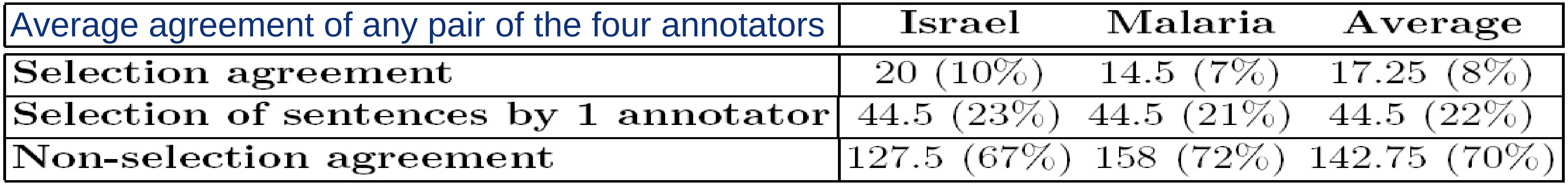
- Four annotators were used in the sentence selection process.

- Agreement from two different points of view:
  - the relative agreement of all four annotators;
  - the average agreement of any pair of the four annotators.

Inter-annotator agreement



Relative agreement among all 4 annotators	Israel	Malaria	Average
Selection agreement of all annotators	10 (5%)	6 (3%)	8 (4%)
Selection agreement of 3 annotators	11 (6%)	10 (5%)	10.5 (5%)
Selection agreement of 2 annotators	27 (14%)	21 (10%)	24 (12%)
Selection by only 1 annotator	42 (22%)	51 (23%)	46.5 (23%)
Non-selection agreement of all annotators	102 (53%)	129 (59%)	115.5 (56%)



Average agreement of any pair of the four annotators	Israel	Malaria	Average
Selection agreement	20 (10%)	14.5 (7%)	17.25 (8%)
Selection of sentences by 1 annotator	44.5 (23%)	44.5 (21%)	44.5 (22%)
Non-selection agreement	127.5 (67%)	158 (72%)	142.75 (70%)

- Relative:
  - a steeper pyramid of agreements with a smaller top;
  - fine-grained discrimination capability due to the higher number of levels.
- Average:
  - a moderate pyramid of agreements with a larger top;
  - coarse discrimination capability.
- “Israel” cluster more compact than the “Malaria” one.

Projecting the summarisation gold-standard to other languages

- For each cluster of documents, given:
  - the selected sentences in English;
  - sentence alignment information for the parallel text collection

  the gold standard of one language can be projected to all other languages.

- The more languages in the parallel corpus, the more time can be saved.

- Problems with unbalanced sentence alignment:
  - One-to-two

  - Two-to-one

Projecting the summarisation gold-standard to other languages

- one-to-two sentence alignment:

  <A-1>***In the absence of special reasons, like a change in sexual partners, there seems to be no reason to prefer the existence of one child to that of the other.***</A-1>

  <B-1>Ohne besondere Gründe, z .</B-1>
  <B-2>B. den Wechsel des Sexualpartners, scheint es keinen Grund zu geben, das Leben eines Kindes dem des anderenvorzuziehen.</B-2>

  - the human selected sentence is added to the gold standard in language A
  - both sentences in language B are added to the gold standard in language B.

Projecting the summarisation gold-standard to other languages

- two-to-one sentence alignment:

<A-1> ***Selecting our children raises more profound ethical problems.*** </A-1>
<A-2>This is not new. <A-2>

<B-1>Le fait de sélectionner nos enfants sur critères soulève des questions éthiques bien plus profondes – ce n’est pas une nouveauté.</B-1>

  - the human selected sentences is added to the gold standard in language A
  - the relative sentence in language B is added to the other gold standard

Outline

- Motivation
- Multilingual parallel evaluation data for summarisation
  - Corpus preparation
  - Human annotation/sentence selection of English documents
  - Automatic projection to all other languages
- ***Methodology for automatic summary evaluation***
  - Comparison across languages
- Released Data
- Conclusion

Using different degrees of inter-annotator agreement

- Idea:
  - use the inter-annotator agreement to rank the selected sentences for each cluster;
- Each sentence is associated to a score: 0 – 4
  - number of annotators that have selected that sentence.
- Better performance of the summariser if:
  - the automatically selected sentences were manually selected by all or most of the annotators.

Summary length unit: word or sentence?

- Most summarisation tasks require the system to produce summaries of a certain length.

- Evaluate several numbers of selected sentences rather than summary lengths:
  - annotators are free to select as many sentences as they think useful.

- Use our produced summaries for summary length comparisons:
  - first select high-ranking annotated sentences;
  - fill the remaining summary space with a relatively high-ranking summary sentence.

Methodology for comparing system summaries

- How to compare automatic summaries against the model summaries produced by annotators?

- The proposed scores are:
  - Weighted Model;
  - Binary Model.

- Automatic summaries created using three different techniques:
  - LSA: an in-house summariser based on LSA technology;
  - Random summariser;
  - Lead: summariser selects the first *k* sentences from each article.

- Report results of summaries with 5, 10, and 15 sentences for all 7 languages.

Weighted Model

- Each human-selected sentence is associated to a model summary weight:
  - agreement of all annotators: a value from 4 to 0.

- For each sentence in the automatically generated summary, the model summary weight was added to the summary score.

- Overall score is computed normalizing the summary score by the maximum reachable score.

$$score_w(Summary) = \frac{\sum_{s \in Summary} msw(s)}{\sum_{s \in Summary} \# annotators}$$

- e.g.
  - 4-0 summary: first set contains one sentence selected by all the annotators and one that is not selected at all: $score_w(sum_1) = {}^{4+0}\!/_{4+4} = 0.5$

Weighted Model

- Results of summaries with 5, 10, and 15 sentences:

	Rnd	Lead	LSA
ar	20%	33%	40%
cz	19%	33%	45%
de	20%	33%	40%
en	19%	33%	38%
es	19%	33%	33%
fr	20%	33%	45%
ru	21%	33%	45%
AVG	20%	33%	41%

	Rnd	Lead	LSA
ar	21%	27%	44%
cz	20%	26%	39%
de	21%	21%	40%
en	20%	28%	42%
es	20%	30%	41%
fr	21%	26%	41%
ru	21%	27%	36%
AVG	21%	27%	40%

	Rnd	Lead	LSA
ar	23%	28%	45%
cz	22%	28%	43%
de	23%	26%	37%
en	22%	28%	39%
es	22%	27%	35%
fr	22%	28%	42%
ru	23%	28%	45%
AVG	22%	28%	42%

- The performance differs from language to language

- Results confirm the need for multilingual summarization evaluation.

Weighted Model

- Weighted Model is not highly discriminative.

- Two summarisers select two sets of sentences:
  - 4-0: first set contains one sentence selected by all the annotators and one that is not selected at all:

$$score_w(sum_1) = {4+0} \big/ {4+4} = 0.5$$

  - 2-2: second set contains two sentences that were annotated by only two annotators:

$$score_w(sum_2) = {2+2} \big/ {4+4} = 0.5$$

- Would a human being prefer the first or the second summary?

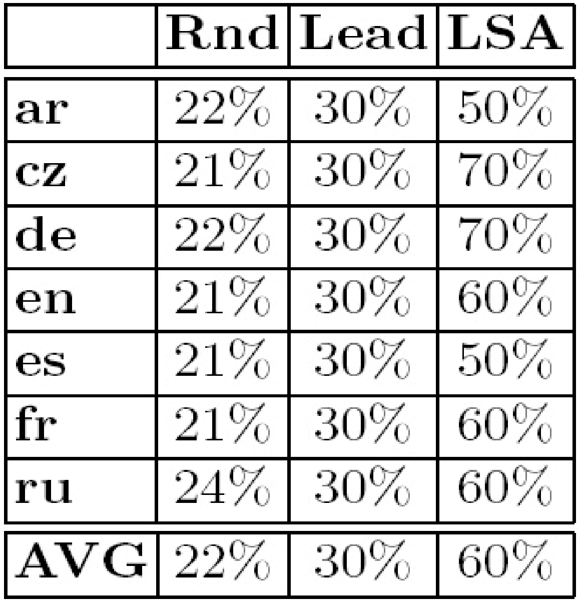
- Are the sentences at the top level two times more important than those selected by two annotators?

Binary Model

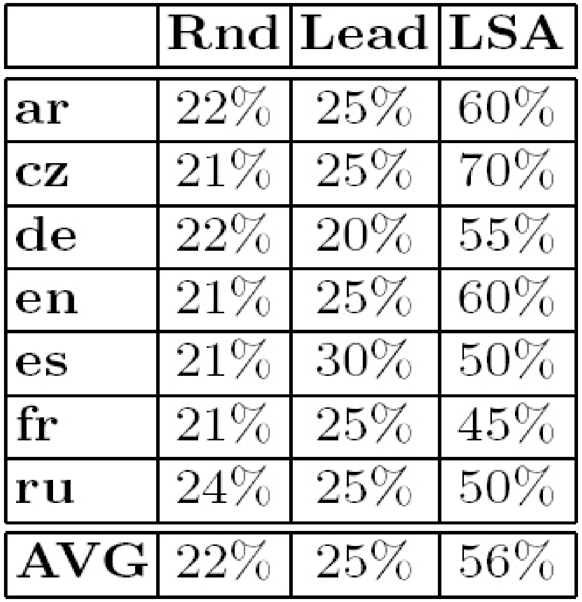
- A more compact sentence scoring approach:
  - a sentence was found important if it was selected by at least two annotators (binary model).
- For each summary:
  - computed the intersection of sentences selected by the summariser with those selected by at least two annotators.
- Overall score is computed as the number of sentences in the intersection divided by the number of sentences in the system summary.

Binary Model

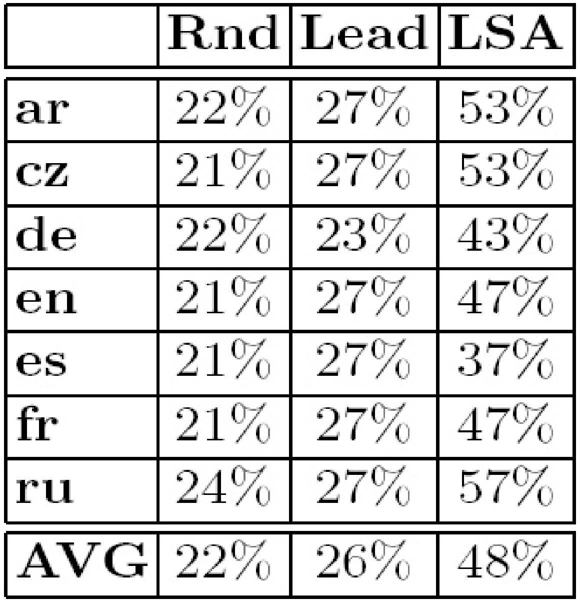
- Results of summaries with 5, 10, and 15 sentences:



	Rnd	Lead	LSA
ar	22%	30%	50%
cz	21%	30%	70%
de	22%	30%	70%
en	21%	30%	60%
es	21%	30%	50%
fr	21%	30%	60%
ru	24%	30%	60%
AVG	22%	30%	60%



	Rnd	Lead	LSA
ar	22%	25%	60%
cz	21%	25%	70%
de	22%	20%	55%
en	21%	25%	60%
es	21%	30%	50%
fr	21%	25%	45%
ru	24%	25%	50%
AVG	22%	25%	56%



	Rnd	Lead	LSA
ar	22%	27%	53%
cz	21%	27%	53%
de	22%	23%	43%
en	21%	27%	47%
es	21%	27%	37%
fr	21%	27%	47%
ru	24%	27%	57%
AVG	22%	26%	48%

- Higher score to the summariser that selects more sentences chosen by more annotators rather than unimportant sentences.
  - gap in performance between LSA and Lead summarisers increases compared to the weighted model.

Binary Model

- On the previous example:
  - 4-0:

$$score_b(sum_1) = {}^{1+0}\!/_{1+1} = 0.5$$

  - 2-2:

$$score_b(sum_2) = {}^{1+1}\!/_{1+1} = 1$$

- Choice of the best set is arbitrary.
- But the binary model disambiguates it in favour of the two-two selection.

Outline

- Motivation
- Multilingual parallel evaluation data for summarisation
  - Corpus preparation
  - Human annotation/sentence selection of English documents
  - Automatic projection to all other languages
- ***Methodology for automatic summary evaluation***
  - ***Comparison across languages***
- Released Data
- Conclusion

Comparison across languages

ar	cz	de	en	es	fr	ru		AVG
1	0.35	*0.28*	0.31	0.34	0.37	0.31	**ar**	0.33
0.35	1	0.37	**0.43**	0.33	0.35	**0.41**	**cz**	0.36
0.28	0.37	1	**0.41**	0.3	0.34	*0.27*	**de**	0.34
0.31	**0.43**	**0.41**	1	**0.43**	**0.41**	0.35	**en**	0.39
0.34	0.33	0.3	**0.43**	1	0.34	*0.28*	**es**	0.34
0.37	0.35	0.34	**0.41**	0.34	1	*0.27*	**fr**	0.36
0.31	**0.41**	*0.27*	0.35	*0.28*	*0.27*	1	**ru**	0.32

- **Bold**: high agreement (> 40%)
- ***Italic***: low agreement (< 30%)

- Percentage of number of sentences shared by the LSA summaries across languages and clusters.
- Quite low agreement, also using statistical summarizer, confirms the need for multilingual summarization evaluation.
- This analysis was not possible before due to the lack of multilingual parallel evaluation data.

Outline

- Motivation
- Multilingual parallel evaluation data for summarisation
  - Corpus preparation
  - Human annotation/sentence selection of English documents
  - Automatic projection to all other languages
- Methodology for automatic summary evaluation
  - Comparison across languages
- ***Released Data***
- Conclusion

Released Data

- Data is available here: http://langtech.jrc.ec.europa.eu/JRC_Resources.html
- For each cluster of documents, we have:
  - One “alignment” file per language
  - One “annotation” file
  - One “data” /“data-annotated” file per language
- “Alignment” file

```
<alignment cid=“Genetic" lang1="English" lang2="French">
    <document did1=“genetic1“ did2=“genetic1">
      <link type="1:1" xtargets="1;1"/>
      …
```

  - cid: cluster id
  - did: document id
  - type: type of alignment
  - xtargets: sentence ids that are involved in the alignment

Released Data

- “Annotation” file

```
<cluster cid=“Genetic">
        <document did=“genetic1">
                <annotation annotators="B D" sid="11"/>
                <annotation annotators="A B D" sid="16"/>
                …
```

  - annotators= ids of the annotators who selected that particular sentence in the English document

- “Data”/ “Data-Annotated” file

```
<cluster cid=“Genetic" lang="English">
    <document did=“genetic1" url="http://www.project-syndicate.org/.../duve1/English">
      <s sid="1“ annotators="B D" >The Origin of Life</s>
      …
```

  - sid: sentence id

Conclusion

- We propose:
  - a ***semi-automatic approach to generate corpora*** for research on multilingual summarisation taking advantage of the parallelism among documents in different languages;
  - an ***evaluation score*** based on different degrees of inter-annotator agreement between human annotators;
  - ***comparison*** of the performance of automatic summarisers ***on seven different languages***.
- Our evaluation method can be applied to evaluate other text mining tools such as information extraction systems.
- The produced data are available for download
  - Thanks to Project Syndicate that gave us the right to use and distribute the data for research purposes.

Thanks a lot for your attention.

Using Parallel corpora for Multilingual (Multi-Document) summarisation Evaluation

Marco Turchi, Josef Steinberger, Mijail Kabadjov,
and Ralf Steinberger
European Commission, Joint Research Centre
Institute for the Protection and Security of the Citizen,
Optima Action
{Name.Surname}@jrc.ec.europa.eu