AllSummarizer system at MultiLing 2015: Multilingual single and multi-document summarization

Abdelkrime Aries, Djamel Eddine Zegour and Khaled Walid Hidouci

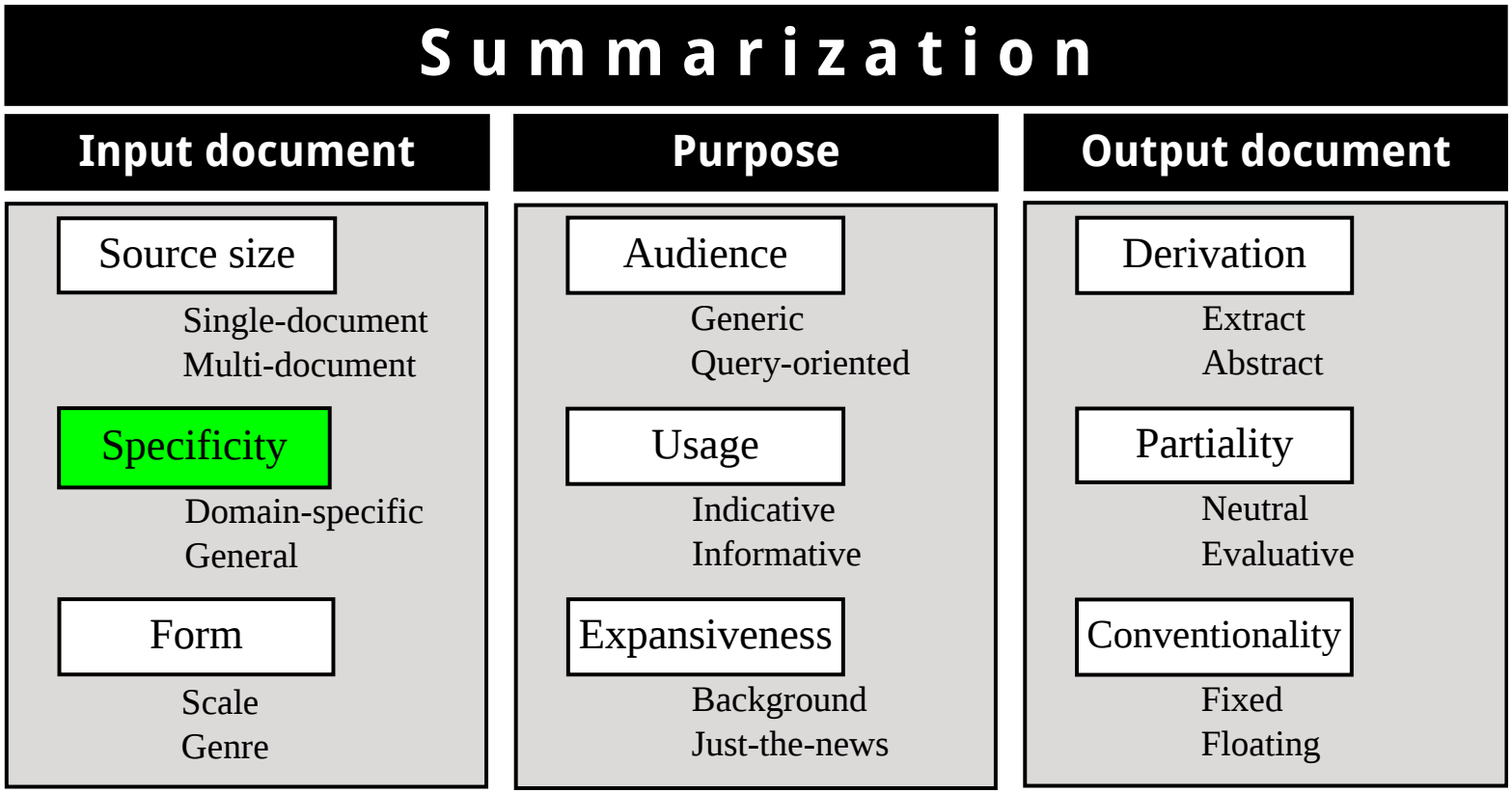
École nationale Supérieure d'Informatique (ESI ex. INI)
Oued-Smar, Algiers, Algeria

September 03, 2015

Introduction
Why summarize?

Introduction
Summarization classification

Following [Hovy and Lin, 1998, Sparck Jones, 1999]:



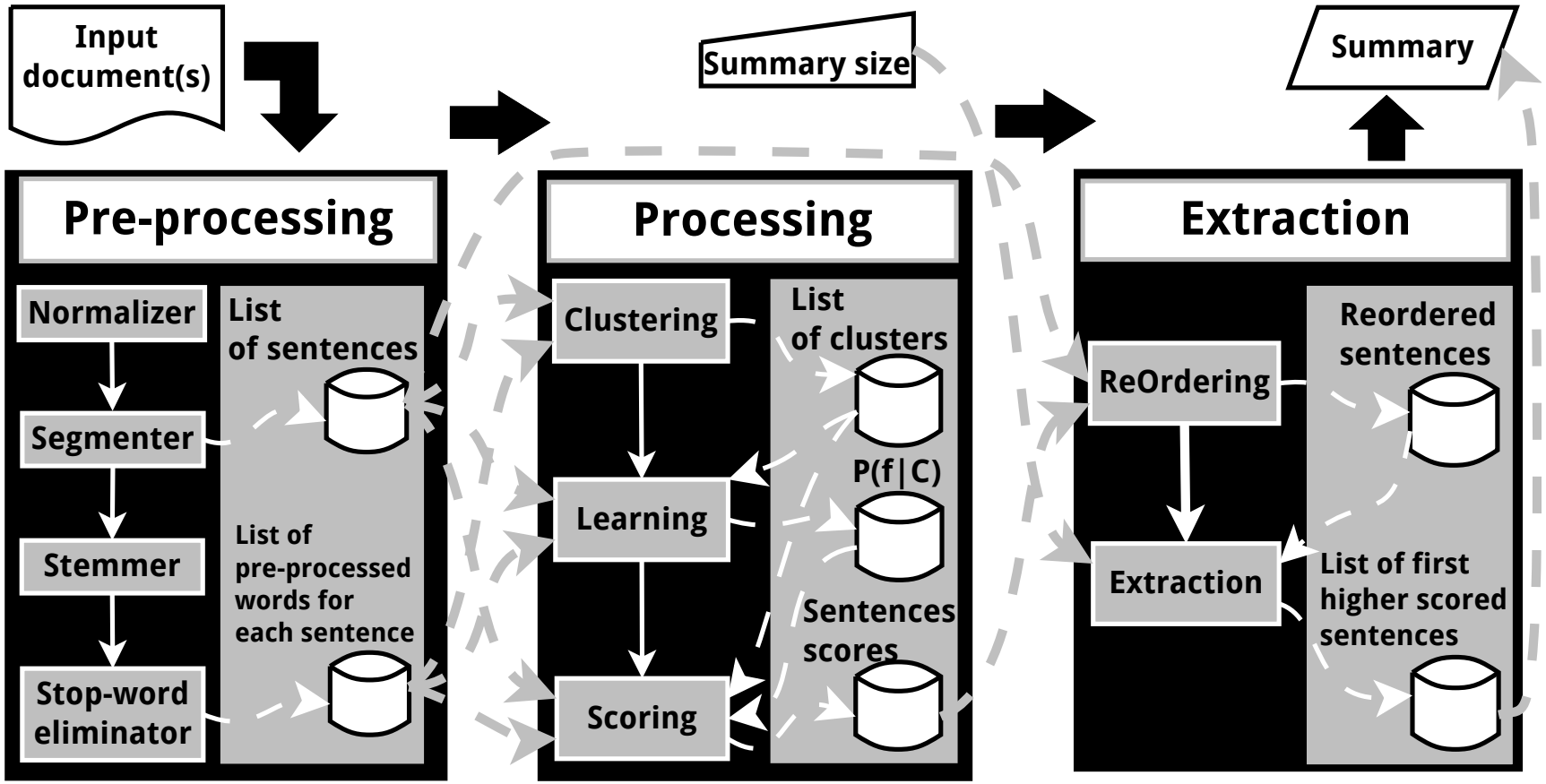
Introduction
Multilingual systems

- Process more than one language.
- Language independent application:
  - Fully independent
  - Partial independent

AllSummarizer system

https://github.com/kariminf/AllSummarizer

System architecture



Preprocessing

Task	Tools	Languages
Sentence segmentation	openNLP	Nl, En, De, It, Pt, Th
	JHazm	Fa
	Regex	The remaining
Words tokenization	openNlp	Nl, En, De, It, Pt, Th
	Lucene	Zh, Ja
	Regex	The remaining
Stemming	Shereen Khoja	Ar
	JHazm	Fa
	HebMorph	He
	Lucene	Bg, Cs, El, Hi, Id, Ja, No
	Snowball	Eu, Ca, Nl, En (Porter), Fi, Fr, De, Hu, It, Pt, Ro, Ru, Es, Sv, Tr
	/	The remaining

Processing
Our idea

Topics of a text:

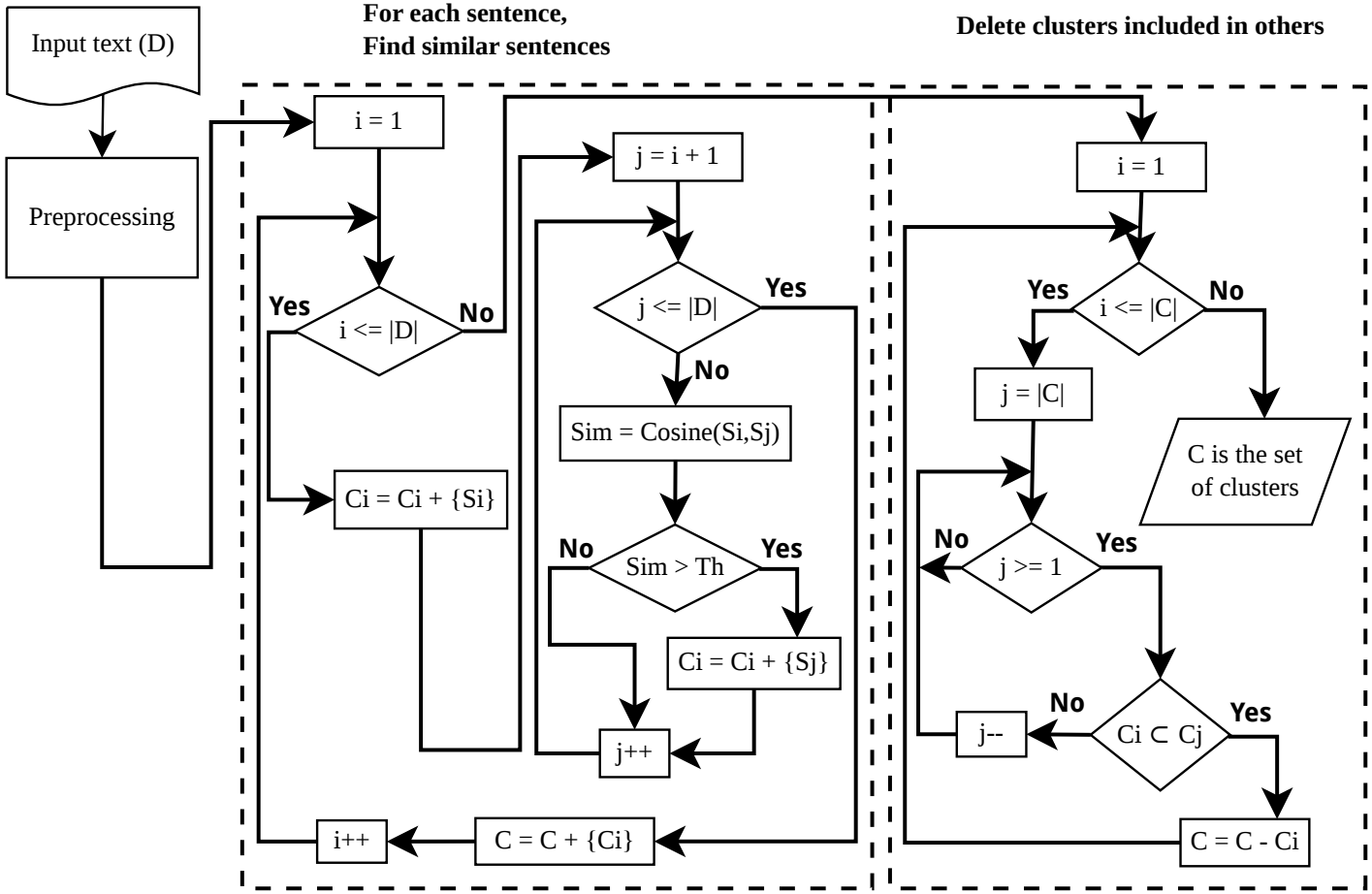
- A text is composed of many topics.
- A sentence can express more than one.

A summary:

- Can represent the content of the input text, and therefore the topics in it.
- Have the most probable sentences to represent all topics.

$\Longrightarrow$ We have to find a method to quantify how much a sentence can represent each and all topics.

Processing
Clustering



Processing
Training

$$P_f(f = \phi | c_j) = \frac{|\phi \in c_j|}{\sum_{c_l \in C} |\phi' \in c_l|}$$

f: feature, ϕ: observation of f, C: set of clusters.

$f \in$

- unigram term frequency (TFU)
- bigram term frequency (TFB)
- sentence position (Pos)
- sentence length (Rleng, PLeng)

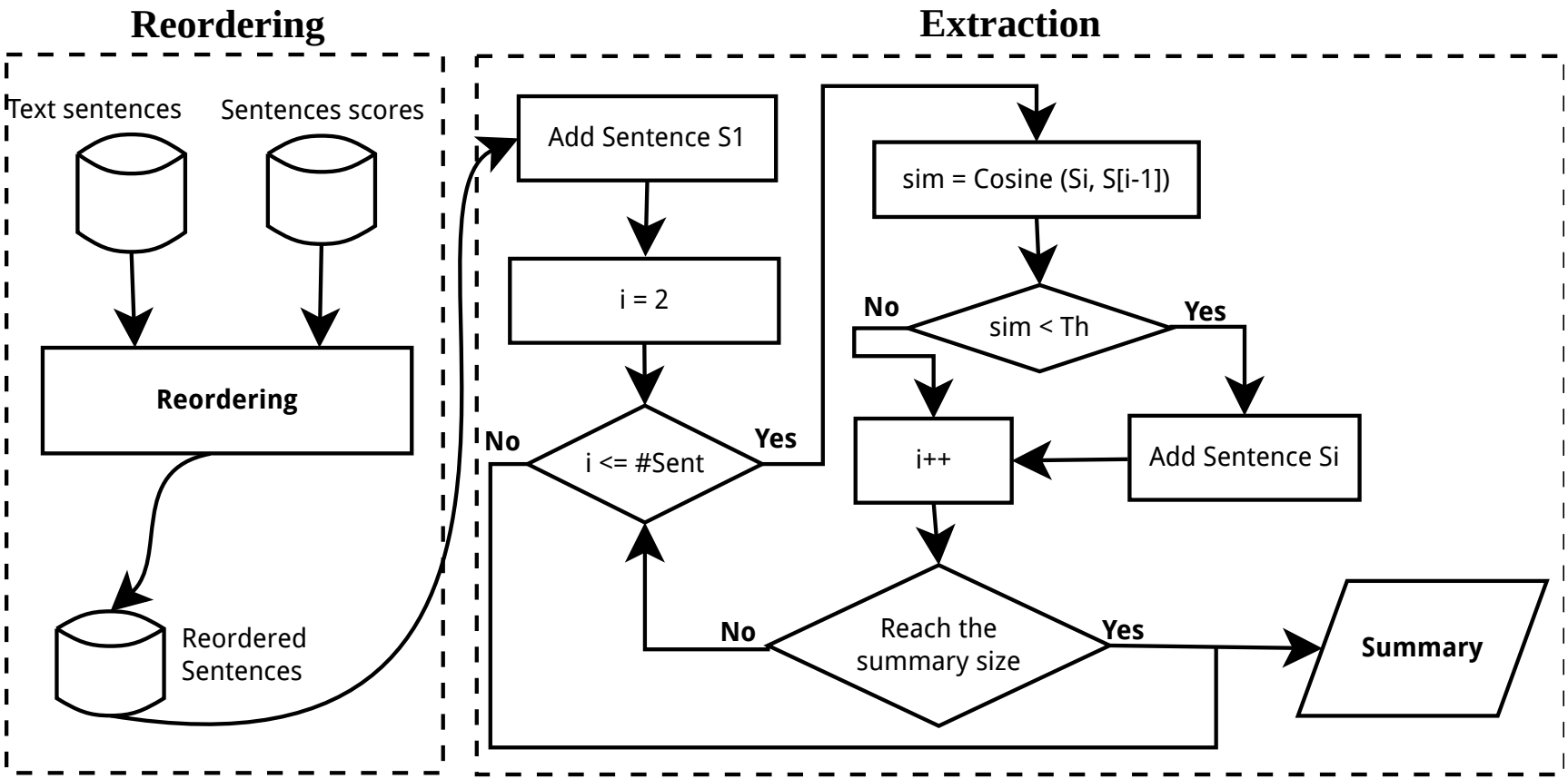
Processing
Scoring

$$Score(s_i, c_j, f_k) = 1 + \sum_{\phi \in s_i} P(f_k = \phi | s_i \in c_j)$$

$$Score(s_i, \bigcap_j c_j, F) = \prod_j \prod_k Score(s_i, c_j, f_k)$$

s: sentence, c: cluster, f: feature, F: set of used features, ϕ: observation of f.

Extraction


Reordering
Extraction
Text sentences
Sentences scores
Reordering
Reordered Sentences
Add Sentence S1
i = 2
No
i <= #Sent
Yes
sim = Cosine (Si, S[i-1])
No
sim < Th
Yes
i++
Add Sentence Si
No
Reach the summary size
Yes
Summary

Experiments

- Parameters estimation.
- Evaluation (Testing).

Parameters estimation

Threshold: Statistic measures

- The median
- The mean
- The mode: lower mode and higher mode.
- The variance
- $sDn = \frac{\sum |s|}{|D| * n}$
- $Dsn = \frac{|D|}{n * \sum |s|}$
- $Ds = \frac{|D|}{\sum |s|}$

$|s|$: number of different terms in a sentence s. $|D|$: number of different terms in the document D. n: number of sentences in this document.

Parameters estimation
Selection process

MMS task training - English:

		TFU-TFB-Pos-RLeng	TFU-TFB-Pos-PLeng	TFU-TFB-RLeng-PLeng	TFU-Pos-RLeng-PLeng	TFB-Pos-RLeng-PLeng	TFU-TFB-Pos-RLeng-PLeng
M001	median	0.0909	0.1105	0.1259	0.1273	0.1385	0.0951
	sDn	0.0783	0.0951	0.0895	0.1385	0.0951	0.1203
	Lmode	0.1147	0.0937	0.1301	0.1497	0.1245	0.0923
	Hmode	0.1147	0.0937	0.1301	0.1497	0.1245	0.0923
	mean	0.0909	0.0909	0.1189	0.0923	0.1063	0.1357
	variance	0.0783	0.0951	0.0895	0.1385	0.0951	0.1203
	Ds	0.1119	0.1119	0.1063	0.1119	0.0531	0.1119
	Dsn	0.0783	0.0951	0.0895	0.1385	0.0951	0.1203
...							
AVG	median	0.0105	0.0108	0.0112	0.0109	0.0122	0.0102
	sDn	0.0075	0.0095	0.0111	0.0110	0.0093	0.0106
	Lmode	0.0106	0.0099	0.0115	0.0133	**0.0133**	0.0100
	Hmode	0.0125	0.0095	0.0115	0.0125	0.0114	0.0100
	mean	0.0109	0.0089	0.0120	0.0097	0.0117	0.0133
	variance	0.0075	0.0095	0.0111	0.0110	0.0093	0.0106
	Ds	0.0091	0.0086	0.0099	0.0100	0.0100	0.0088
	Dsn	0.0075	0.0095	0.0111	0.0110	0.0093	0.0106

Parameters estimation

Selected parameters

Lang	Single document (MSS)		Multidocument (MMS)	
	Th	Features	Th	Features
Ar	Ds	TFB, Pos, PLeng	Ds	TFB, Pos, RLeng, PLeng
Cs	HMode	TFU, TFB, Pos, PLeng	Ds	TFB, Pos, PLeng
El	Median	TFU, TFB, Pos, RLeng, PLeng	LMode	TFB, RLeng
En	Median	TFU, Pos, RLeng, PLeng	LMode	TFB, Pos, RLeng, PLeng
Es	sDn	TFB, PLeng	Ds	TFB, PLeng
Fr	Median	TFB, Pos, RLeng	Mean	TFU, TFB, Pos, PLeng
He	Ds	TFB, PLeng	Median	TFB, RLeng, PLeng
Hi	/	/	Ds	TFB, Pos, RLeng, PLeng
Ro	HMode	TFB, RLeng, PLeng	sDn	TFB, Pos, PLeng
Zh	HMode	TFB, RLeng, PLeng	sDn	TFU, Pos, RLeng, PLeng

System evaluation
Comparing criteria

Let AS = AllSummarizer
S = other system participated with n languages

$$AVG_S = \frac{\sum_{i=1}^{n} Score_S(L_i)}{n}$$

$$AVG_{AS} = \frac{\sum_{i=1}^{n} Score_{AS}(L_i)}{n}$$

Relative improvement (RI):

$$RI = \frac{AVG_{AS} - AVG_S}{AVG_S}$$

System evaluation
Single document (MSS task)

Methods	Our method improvement %				
	R-1	R-2	R-3	R-4	R-SU4
BGU-SCE-M (ar, en, he)	-09.19	-14.02	-19.39	-25.12	-11.07
EXB (all 38)	-07.64	-10.55	-09.86	-07.92	-10.63
CCS (all 38)	-07.33	-13.24	-10.95	-03.04	-07.40
BGU-SCE-P (ar, en, he)	-04.33	-01.63	-02.69	-06.16	-01.89
UA-DLSI (en, de, es)	+02.12	+06.25	+13.86	+17.15	+05.62
NTNU (en, zh)	+06.44	+07.06	+11.50	+21.81	+05.74
Oracles (all 38) [TopLine]	-31.64	-49.00	-63.80	-72.91	-36.77
Lead (all 38) [BaseLine]	+02.39	+08.67	+08.20	+04.02	+05.82

System evaluation
Multidocument (MMS task)

SysID	Our method improvement %		
	AutoSummENG	MeMoG	NPowER
UJF-Grenoble (fr, en, el)	-08.87	-14.55	-03.62
UWB (all 10)	-22.56	-22.66	-07.54
ExB (all 10)	-09.44	-09.16	-02.80
IDA-OCCAMS (all 10)	-17.11	-17.68	-05.53
GiauUngVan (- zh, ro, es)	-16.43	-19.40	-05.68
SCE-Poly (ar, en, he)	-05.72	-03.35	-01.46
BUPT-CIST (all 10)	+10.67	+11.53	+02.85
BGU-MUSE (ar, en ,he)	+05.67	+06.92	+01.74
NCSR/SCIFY-NewSumRerank (- zh)	+01.53	-01.25	+00.13
AllSummazer (MSS param) (all 10)	+01.98	+02.35	+00.58

Conclusion and Perspectives

These experiments have shown that:

- Can't use the same parameters (Th & $\overrightarrow{F}$) to every language.
- Estimating the parameters using the average doesn't give the best score.
- Fair results.

For future:

- Estimate the parameters based on the input text using statistical criteria.
- Investigate the effect of preprocessing and clustering on the result.
- Readability in a multilingual context.

Bibliography


Hovy, E. and Lin, C.-Y. (1998).
Automated text summarization and the SUMMARIST system.
In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*, pages 197–214. Association for Computational Linguistics.


Sparck Jones, K. (1999).
Automatic summarising: factors and directions.
In *Advances in automatic text summarisation*. Cambridge MA: MIT Press.

Thank you for your attention

Questions