KTH Data- och
systemvetenskap

Swedish extract corpus

Hercules Dalianis

Department of Computer
and System Sciences

Stockholm University and KTH

Email:hercules@kth.se

Two experiment

- KTH extract tool
- Students, colleagues, friends
- Språkkonsulter (Students of professional writing)

KTH Data- och systemvetenskap

Collecting extract corpus

- 10 news text from Svenska Dagbladet
- Experiment fall 2003
- 28 informants students, colleagues, friends,
- We obtained 238 Swedish extract
- average length of 34 percent of original texts
- 61 percent overlap at average length

SweSum and extract corpus

- Comparing performance of SweSum with Swedish extract corpus
- Majority vote extract at average length ideal extract
- SweSum - with input of original text and average length parameters

Quality Correlation

					Comparison with SweSum summary				
Swedish	Extracts	Average	Overlap	Overlap	Overlap at			truncated	truncated
Extracts	No of	extract	of	at average	sentence	word	word	word	word
Filename	extracts	length	all votes	length	level	level	frequency	level	frequency
text001.htm	28	37%	36%	60%	80%	81%	75%	84%	74%
text002.htm	19	27%	33%	60%	46%	53%	42%	54%	42%
text003.htm	22	30%	33%	57%	83%	94%	90%	94%	90%
text004.htm	16	31%	36%	70%	78%	84%	79%	85%	79%
text005.htm	24	33%	34%	59%	59%	67%	63%	66%	61%
text006.htm	29	32%	33%	62%	44%	64%	53%	65%	53%
text007.htm	26	39%	35%	60%	75%	70%	65%	69%	64%
text008.htm	22	37%	35%	62%	36%	63%	56%	65%	58%
text009.htm	24	32%	33%	56%	59%	79%	70%	82%	70%
text010.htm	28	41%	34%	65%	57%	75%	66%	75%	66%
Total/Average	238	34%	34%	61%	62%	73%	66%	74%	66%

61 percent average overlap for corpus

62 percent average overlap for SweSum

Some texts up to 80 percent overlap

Språkkonsulter

- 5 texts
- 63 extracts
- 15 informants

Quality Correlation

Språkkonsulter					Comparison with SweSum summary				
Swedish 2	Extracts	Average	Overlap	Overlap	Overlap at			truncated	truncated
Extracts	No of	extract	of	at average	sentence	word	word	word	word
Filename	extracts	length	all votes	length	level	level	frequency	level	frequency
text001.htm	12	33%	48%	79%	50%	62%	52%	63%	51%
text002.htm	11	22%	39%	69%	34%	43%	33%	44%	33%
text003.htm	11	34%	38%	69%	63%	70%	59%	71%	59%
text004.htm	15	32%	42%	80%	29%	31%	23%	31%	22%
text005.htm	14	32%	42%	67%	57%	70%	61%	71%	60%
Total/Average	63	31%	42%	73%	47%	55%	46%	56%	45%

- What about looking at one extract at a time and not at majority votes?