Researching Responsible Natural Language Processing

Session 15: Scientific Writing: Graphs, Diagrams, Tables

Frank Keller

6 November 2024

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Overview

Graphs

Diagrams

Captions

Tables

Reading: Zobel (2014), ch. 11.

Please also look at Alley's web site, which has a lot of videos and additional materials: https://www.craftofscientificwriting.com/

Overview

Visual material is an important part of a paper:

- diagrams illustrate complex ideas, processes, or models
- graphs show trends or relationships in data
- tables present results or regularities in data
- textual panels present algorithms or mathematical formulas

Such materials attracts the reader's attention; some readers will only look at figures and tables, they will not read (all of) the text.

Graphs

Graphs

- graphs can make behaviors and trends obvious that a hard to discern from a table
- keep graphs simple, avoid both clutter and unnecessary whitespace
- for elements such as secondary ticks, legends, gridlines, boxes, ask if you really need them
- use the same fonts in graphs and tables as in the main text
- sometimes logarithmic axes are appropriate
- a table of results can often be represented as a bar graph
- if you use multiple graphs to display the same quantity, use the same axis (same range) in all of them

Graphs

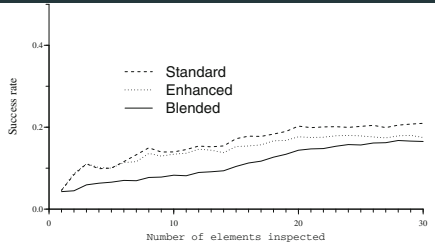

FIGURE 7. *Success rate as the number of inspected items is increased. It is clear that blending is not effective.*

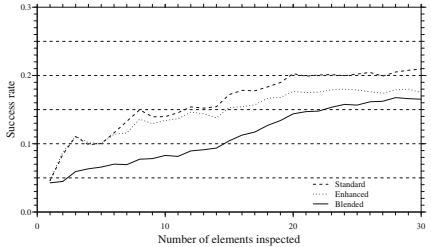

FIGURE 7. *Success rate as the number of inspected items is increased. It is clear that blending is not effective.*

Graphs

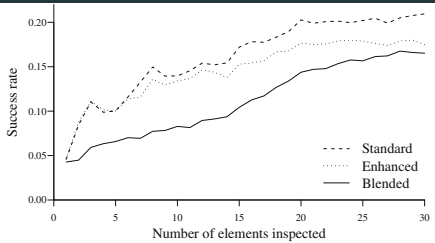

FIGURE 7. *Success rate as the number of inspected items is increased. It is clear that blending is not effective.*

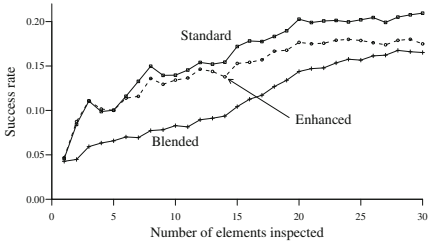

FIGURE 7. *Success rate as the number of inspected items is increased. It is clear that blending is not effective.*

Graphs

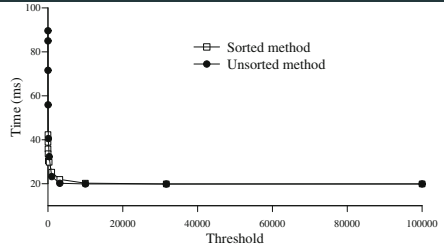

FIGURE 6. *Evaluation time (in milliseconds) for bulk insertion methods as threshold is varied.*

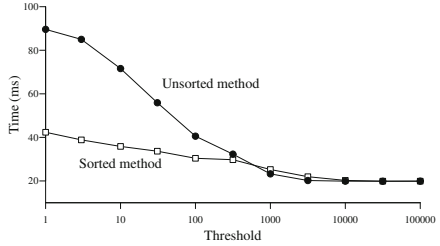

FIGURE 6. *Evaluation time (in milliseconds) for bulk insertion methods as threshold is varied.*

Graphs

Data set	Method	
	A	B
Small, random	11.5	11.6
Large, random	27.9	17.1
Small, clustered	9.7	8.2
Large, clustered	24.0	13.5
All documents	49.4	60.1
First 1000	21.1	35.4
Last 1000	1.0	5.5

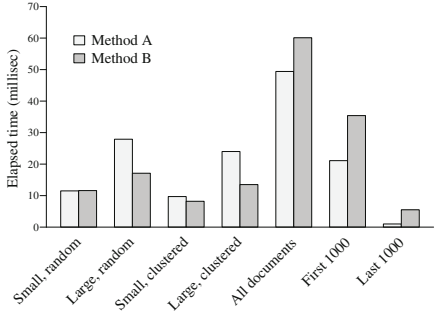

FIGURE 2. *Elapsed time (milliseconds) for methods A and B applied to data sets 1–7.*

Diagrams

Diagrams

- diagrams show architectures, structures, processes, relationships, or states
- typically, the diagram should just show **one** of the things; an attempt to combine them often makes the diagram less clear
- it's a good idea to sketch the diagram by hand first, check layout, proportions, use of space, sizes of elements
- focus on the concept being illustrated, avoid clutter and unnecessary detail
- use pictorial elements consistently (arrows or boxes of the same kind always have the same meaning, etc.)
- don't expect to get it right first time, revise your diagrams as you would revise your text

Diagrams

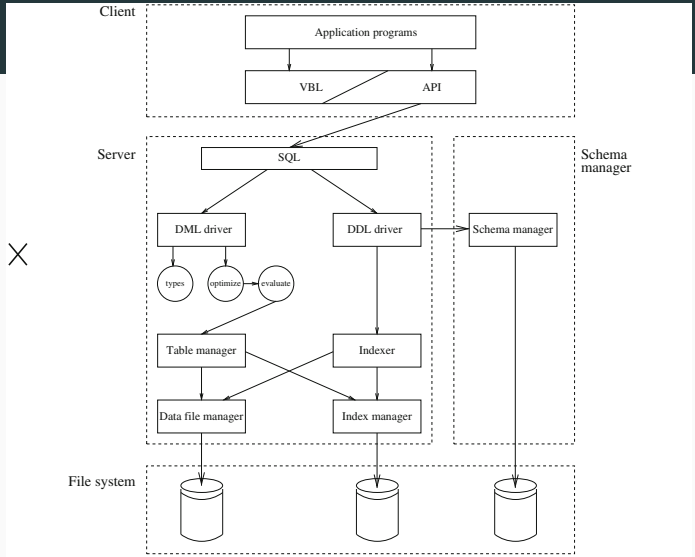

FIGURE 1.3. *System architecture, showing the relationship between the major components. Each component is an independent process. Note the lack of a single interface to the file system.*

Diagrams

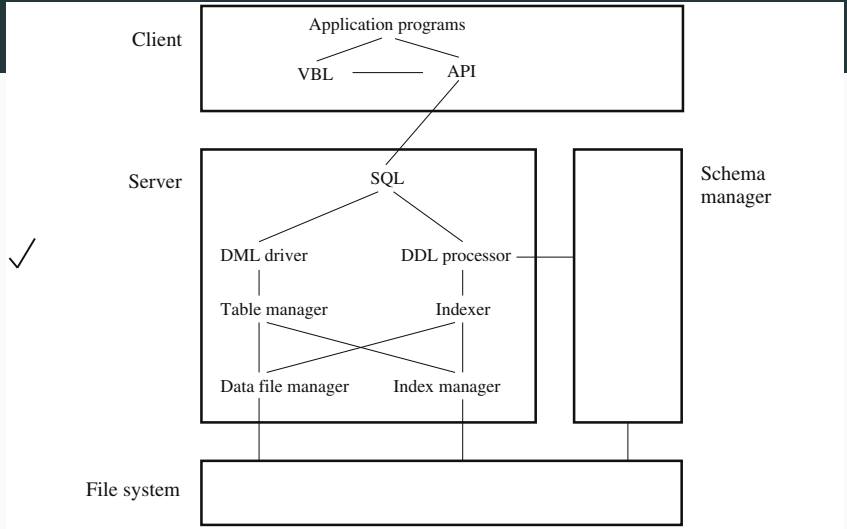

FIGURE 1.3. *System architecture, showing the relationship between the major components. Each component is an independent process. Note the lack of a single interface to the file system.*

Diagrams

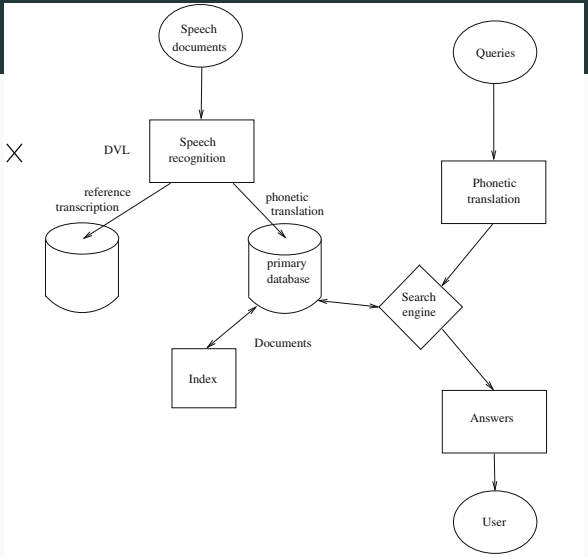

FIGURE 7. *The* QUIRK *system for matching written queries to speech. Each input document is translated into a string of phonemes and then stored. Queries are also translated into phonemes, which can be matched to the documents. Answers are returned to the user.*

Diagrams

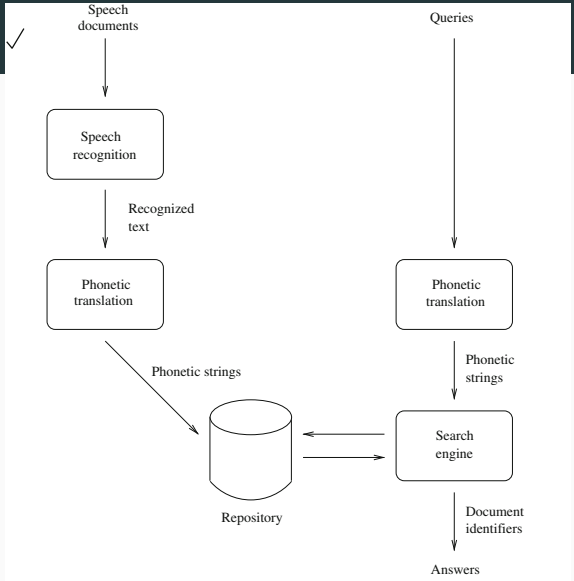

FIGURE 7. *The* QUIRK *system for matching written queries to speech. Each input document is translated into a string of phonemes and then stored. Queries are also translated into phonemes, which can be matched to the documents.*

Captions

Captions

- captions should fully describe the major elements of a figure or table
- together with its caption, the figure or table should be self-contained, i.e., understandable without referring to the text
- captions should assist a reader who's only skimming the paper, or who is going back to re-read parts of a longer paper
- normally, the caption appears above a table, but below a figure
- if you use abbreviations or symbols in a figure or table, then these need to be explained in the caption
- the caption can also contain additional detail that would interrupt the flow of the main text

Over to You

Exercise 1

Let's return to Lee et al. (2024) and Chen et al. (2024), the two papers on explainable multimodal NLP that we look at last time.

The following page show to examples of diagrams from these papers.

- Are the diagrams well-designed?
- Does they have the right level of complexity?
- Are the captions appropriate?

How would you modify the diagrams and captions to improve them?

Exercise 1: Lee et al. (2024)

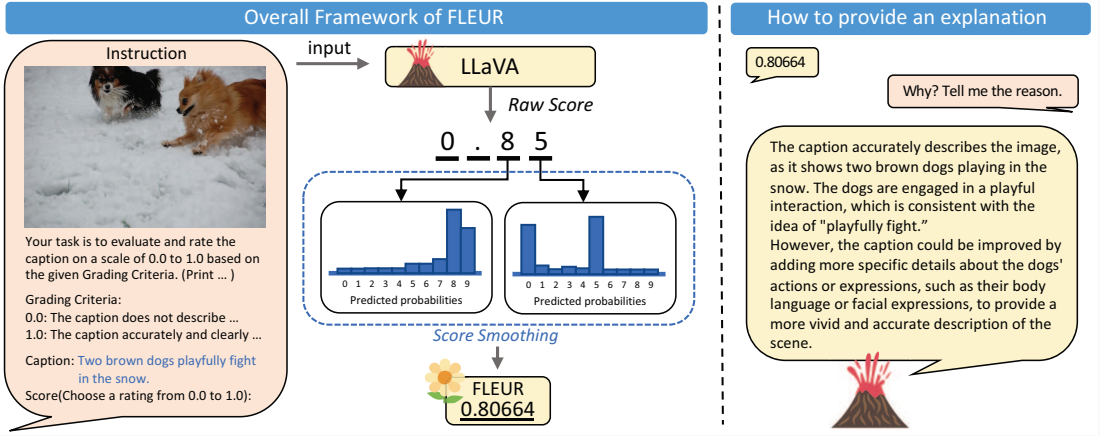

Figure 2: The overall framework of FLEUR. **Left**: When feeding LLaVA with the prompt containing the grading criteria, image, and the candidate caption for evaluation, FLEUR takes a weighted sum of probabilities of tokens (0 to 9) as the final score. **Right**: When prompted by the user for the rationale behind the given score, FLEUR provides explanations in a language understandable to humans.

Exercise 1: Chen et al. (2024)

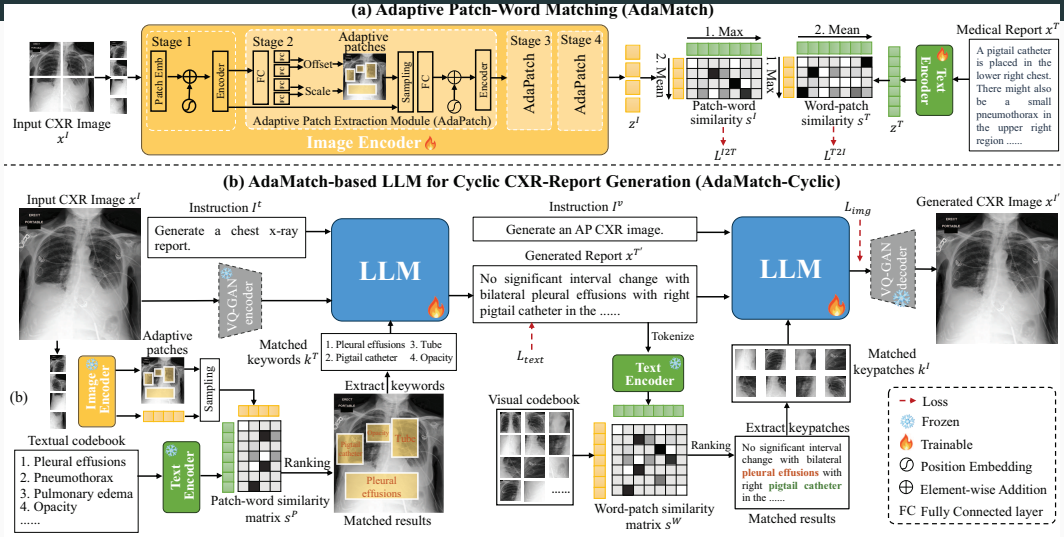

Figure 2: The overview of the proposed methods. (a) Adaptive patch-word Matching (AdaMatch) model. (b) AdaMatch-based bidirectional large language model (LLM) for cyclic CXR-report generation (AdaMatch-Cyclic).

Tables

Tables

- some information cannot be presented easily in graphs or diagrams
- in some cases, the exact numeric values are important
- tables are more suitable than graphs if only a small number of values need to be displayed
- tables can have a hierarchical structure: columns and rows can be partitioned or have internal structure
- the structure needs to be indicated by headings, labels, dividers
- limit the use of horizontal rules; vertical rules should be avoided; tables should contain sufficient whitespace
- don't make a table too big; instead, use two tables or a graph

TABLE 6. *Statistics of text collections used in experiments.*

STATISTICS	SMALL	LARGE
Characters	18,621	1,231,109
Words	2,060	173,145
After stopping	1,200	98,234
Index size	1.31 Kb	109.0 Kb

TABLE 6. *Statistics of text collections used in experiments.*

	Collection	
	Small	Large
File size (Kb)	18.2	1,202.3
Index size (Kb)	1.3	109.0
Number of words	2,060	173,145
— after stopping	1,200	98,234

TABLE 11. *Resources used during compression and indexing. Only the vocabulary is constructed in the first pass; the other structures are built in the second pass.*

Pass	Output	Size Mb	Size %	CPU Hr:Min	Mem Mb
Pass 1:					
Compression	Model	4.2	0.2	2:37	25.6
Inversion	Vocabulary	6.4	0.3	3:02	18.7
Overhead				0:19	2.5
Total		10.6	0.5	5:58	46.8
Pass 2:					
Compression	Text	605.1	29.4	3:27	25.6
	Doc. map	2.8	0.1		
Inversion	Index	132.2	6.4	5:25	162.1
	Index map	2.1	0.1		
	Doc. lens	2.8	0.1		
	Approx. lens	0.7	0.0		
Overhead				0:23	2.5
Total		745.8	36.3	9:15	190.2
Overall		756.4	36.8	15:13	190.2

TABLE 11. *Resources used during compression and indexing. Only the vocabulary is constructed in the first pass; the other structures are built in the second pass.*

Task	Size (Mb)	CPU (Hr:Min)	Memory (Mb)
Pass 1:			
Compression	4.2	2:37	25.6
Inversion	6.4	3:02	18.7
Overhead	—	0:19	2.5
Total	10.6	5:58	46.8
Pass 2:			
Compression	607.9	3:27	25.6
Inversion	137.8	5:25	162.1
Overhead	—	0:23	2.5
Total	745.8	9:15	190.2
Overall	756.4	15:13	190.2

Over to You

Exercise 2

Here are some tables from Lee et al. (2024) and Chen et al. (2024).

- Is the table layout good? How about the use of whitespace?
- Can you decode the hierarchical structure of these tables?
- Should they maybe have used a graph instead?
- Are the captions appropriate?

How would you modify the tables and captions to improve them?

Exercise 2: Lee et al. (2024)

Type	Exp	Metric	Flickr8k		COM	Pascal-50S (Accuracy ↑)				
			EX (τ_c ↑)	CF (τ_b ↑)	(τ_c ↑)	HC	HI	HM	MM	Avg
reference -based		BLEU-4	30.8	16.9	30.6	53.0	92.4	86.7	59.4	72.9
		ROUGE-L	32.3	19.9	32.4	51.5	94.5	92.5	57.7	74.1
		METEOR	41.8	22.2	38.9	56.7	97.6	94.2	63.4	78.0
		CIDEr	43.9	24.6	37.7	53.0	98.0	91.5	64.5	76.8
		SPICE	44.9	24.4	40.3	52.6	93.9	83.6	48.1	69.6
		BERTScore	39.2	22.8	30.1	65.4	96.2	93.3	61.4	79.1
	✓	CLAIR[4]	48.3	–	61.0	52.4	99.5	89.8	73.0	78.7
		TIGEr	49.3	–	45.4	56.0	**99.8**	92.8	74.2	80.7
		ViLBERTScore-F	50.1	–	52.4	49.9	99.6	93.1	75.8	79.6
		RefCLIPScore	53.0	36.4	55.4	64.5	99.6	95.4	72.8	83.1
		RefPAC-S	55.9	37.6	57.3	67.7	99.6	96.0	75.6	84.7
		Polos	**56.4**	37.8	57.6	**70.0**	99.6	97.4	**79.0**	**86.5**
	✓	RefFLEUR (Ours)	51.9	**38.8**	**64.2**	68.0	**99.8**	**98.0**	76.1	85.5
reference -free		CLIPScore	51.2	34.4	53.8	56.5	99.3	96.4	70.4	80.7
		PAC-S	54.3	36.0	55.7	60.6	99.3	96.9	72.9	82.4
		InfoMetIC+[5]	**55.5**	36.6	59.3	–	–	–	–	–
	✓	FLEUR (Ours)	53.0	**38.6**	**63.5**	**61.3**	**99.7**	**97.6**	**74.2**	**83.2**

Table 1: Overall correlation and accuracy comparison with human judgment on Flickr8k-Expert (Flickr8k-EX), Flickr8k-CF, COMPOSITE (COM), and Pascal-50S datasets. Bold indicates the best result in each type. 'Exp' stands for 'explainable' and checkmarks are applied only to the corresponding metrics. FLEUR is the only metric satisfying both explainable and reference-free. All results except for ours are reported results from prior works.

Exercise 2: Chen et al. (2024)

Table 1: Comparison of CXR-to-report generation performance on the MIMIC-CXR and the OpenI datasets.

	MIMIC-CXR						OpenI					
Methods	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
R2Gen	0.3553	0.2232	0.1523	0.1038	0.1412	0.2784	0.3992	0.2407	0.1518	0.0973	0.1390	0.3052
R2GenCMN	0.3719	0.2332	0.1538	0.1053	0.1501	0.2827	0.4091	0.2493	0.1594	0.1045	0.1509	0.3181
Joint-TriNet	0.3585	0.2266	**0.1550**	0.1021	0.1425	0.2788	0.3833	0.2409	0.1598	0.1078	0.1457	0.3293
XProNet	0.3532	0.2212	0.1498	0.1052	0.1415	0.2811	0.4114	0.2502	0.1598	0.1045	0.1457	0.3240
ITHN	0.3623	0.2128	0.1402	0.0992	0.1488	0.2622	0.2661	0.1516	0.0976	0.0663	0.1561	0.2617
M2KT	0.3661	0.2192	0.1465	0.1044	0.1528	0.2673	0.2559	0.1381	0.0819	0.0523	0.1468	0.2439
AdaMatch-Cyclic	**0.3793**	**0.2346**	0.1540	**0.1060**	**0.1625**	**0.2859**	**0.4161**	**0.3002**	**0.2073**	**0.1446**	**0.1621**	**0.3656**

References

Chen, Wenting, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, pages 9494–9509.

Lee, Yebin, Imseong Park, and Myungjoo Kang. 2024. FLEUR: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, pages 3732–3746.

Zobel, Justin. 2014. *Writing for Computer Science*. Springer, New York, NY, 3 edition.