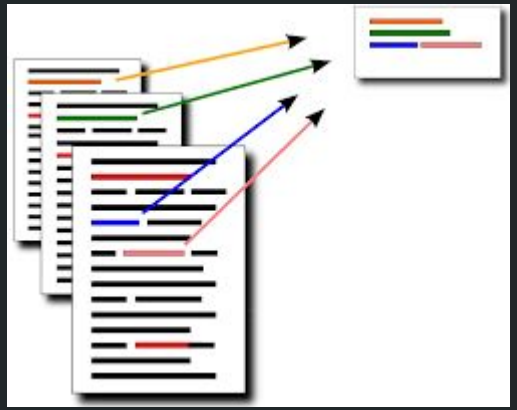
GameWikiSum: a Novel Large Multi-Document Summarization Dataset

Diego Antognini & Boi Faltings

Overview

- Introduction
- Dataset
- Experiments & Results
- Conclusion

Introduction



Introduction : Types of Summarization

- Extractive Summarization
- Abstractive Summarization

Dataset : Dataset Creation

- Metacritic [1]
- Wikipedia [2]

[1] https://www.metacritic.com
[2] https://en.wikipedia.org/wiki/Main_Page

Dataset : Heuristic matching

- Exact title match.
- Removing tags.
- Extension match.

Dataset: Descriptive Statistics

Percentile	20	40	50	60	80	100
Num Documents	2	5	7	10	18	84
Summary Size	139	246	321	419	684	4639
Documents Size	2 536	5 604	7 815	10 634	20 498	249 062
ROUGE-1 recall	67.7	80.7	85.29	88.8	94.1	100.0
ROUGE-2 recall	14.3	23.0	27.4	31.9	41.9	100.0

Table 1: Percentiles for different aspects of GameWikiSum. Size is in number of words. ROUGE scores are computed with a summary given its reviews.

Dataset: Descriptive Statistics

Dataset	Input	Output	# Examples	ROUGE-1 R
*Gigaword (Graff and Cieri, 2003)	10^1	10^1	10^6	78.7
*CNN/DailyMail (Nallapati et al., 2016)	10^2-10^3	10^1	10^5	76.1
DUC 2001-2004[5]	10^3	10^2	10^2	94.4
TAC 2008-2011[6]	10^3	10^2	10^2	95.3
WikiSum (Liu et al., 2018)	10^2-10^6	10^1-10^3	10^6	59.2
GameWikiSum (ours)	10^3-10^5	10^2-10^3	10^4	80.1

Table 2: Sizes and unigram recall of single (marked with *) and multi-document summarization datasets. Recall is computed with reference summaries given the input documents.

Dataset: Descriptive Statistics

Platform	# Games	# Documents	ROUGE-1 R	ROUGE-2 R
PC	3586	8 ± 8	81.18 ± 15.45	27.32 ± 14.52
Wii U	224	10 ± 13	86.47 ± 10.78	34.14 ± 16.03
Nintendo 64	66	8 ± 3	77.46 ± 13.10	21.11 ± 9.37
Dreamcast	83	6 ± 2	66.12 ± 13.73	13.01 ± 6.27
PlayStation	86	4 ± 2	60.95 ± 14.67	10.97 ± 6.47
PlayStation 2	954	13 ± 9	85.93 ± 11.74	30.47 ± 11.89
Game Boy Advance	368	5 ± 4	69.38 ± 17.78	17.23 ± 11.15
GameCube	341	10 ± 7	82.26 ± 12.16	24.95 ± 10.66
Xbox	486	15 ± 9	88.40 ± 9.95	32.31 ± 10.79
DS	679	10 ± 9	85.27 ± 11.77	30.99 ± 13.38
PSP	407	12 ± 9	85.08 ± 13.85	30.71 ± 13.27
Xbox 360	1358	19 ± 14	86.90 ± 14.54	34.93 ± 15.72
PlayStation 3	1128	13 ± 11	84.53 ± 16.27	32.28 ± 15.48
Wii	665	10 ± 10	84.70 ± 14.07	32.18 ± 14.77
iOS	1344	4 ± 3	77.86 ± 15.48	23.39 ± 13.26
Xbox One	817	8 ± 9	83.33 ± 14.53	30.66 ± 15.63
3DS	312	15 ± 14	88.62 ± 12.87	39.75 ± 19.01
PlayStation Vita	337	7 ± 9	80.97 ± 14.50	28.21 ± 16.63
PlayStation 4	1103	14 ± 14	87.42 ± 14.02	37.84 ± 18.00
Switch	308	11 ± 12	89.97 ± 9.64	38.61 ± 15.95
All	14652	11 ± 11	83.19 ± 15.04	29.99 ± 15.48

Table 3: Game distribution over platforms with their average and standard deviation number of input documents and ROUGE scores.

Experiments & Results

Experiments & Results : Evaluation Metrics

- ROUGE-L F1.
- ROUGE-1.
- ROUGE-2.

Experiments & Results : Baselines

- LEAD-k [3].
- TextRank [4].
- LexRank [5].
- SumBasic [6].
- C_SKIP [7].
- SemSenSum [8].
- Conv2Conv [9].
- Transformer [10].
- TransformerLM [11]

[3] Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1073–1083
[4] Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. In Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing, July.
[5] Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22:457–479.
[6] Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization. page 101, Microsoft Research, Redmond, Washington, Tech. Rep.
[7] Rossiello, G., Basile, P., and Semeraro, G. (2017). Centroid-based text summarization through compositionality of word embeddings. In Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, pages 12–21.
[8] Antognini, D. and Faltings, B. (2019). Learning to create sentence semantic relation graphs for multi-document summarization. In Proceedings of the 2nd Workshop on New Frontiers in Summarization, ages 32–41, Hong Kong, China, November. Association for Computational Linguistics.
[9] Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1243–1252. JMLR. org.
[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008.
[11] Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. International Conference on Learning Representations.

Experiments & Results : Results

3.3. Results

Model	R-L	R-1	R-2
LEAD-3	11.45	12.77	2.45
LEAD-5	18.78	19.82	3.42
TextRank	29.30	31.07	4.96
LexRank	29.74	31.26	4.96
SumBasic	30.36	31.82	4.79
C_SKIP	31.66	32.90	5.25
SemSenSum	31.72	35.11	5.56
Conv2Conv*	20.10	19.30	5.20
Transformer*	14.60	16.00	2.80
TransformerLM*	9.52	7.03	1.17

Table 4: Comparison extractive and abstractive (marked with *) models. Reported scores correspond to ROUGE-L F1 score, ROUGE-1 and ROUGE-2 recall respectively.

Conclusion

Thank You