Generating Wikipedia by Summarizing Long Sequences

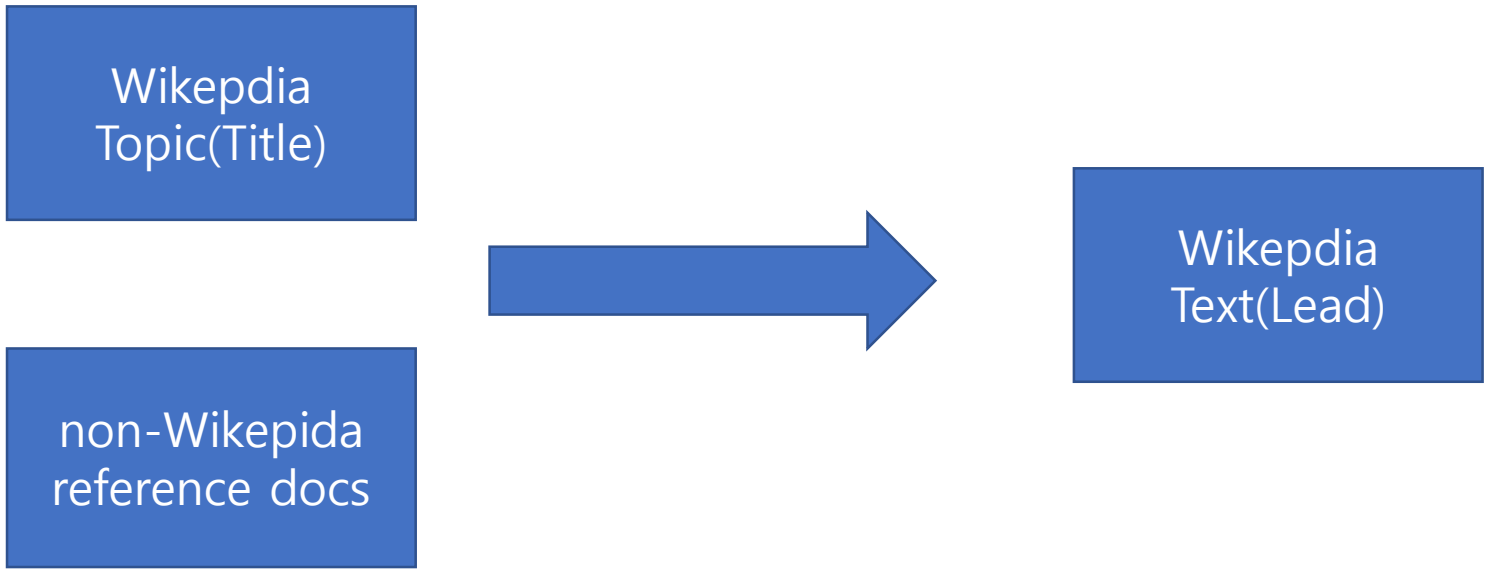
ICLR 2018

Peter J. Liu, Mohammad Saleh
Google Brain

code : https://github.com/tensorflow/tensor2tensor/tree/master/tensor2tensor/data_generators/wikisum

Abstract

- WikiSum Dataset
- Text Generation : Multi Document Summarization 관점에서 접근
- Transformer Decoder Only 처음으로 제안



Summarization

- Extractive Summary : 원문 단어만을 사용해 Summary(Ctrl CV)

ex)

- Abstractive Summary : 원문 외의 단어를 활용해 Summary

ex)

- Multi-document Summary : 여러 document를 한 번에 Summary

Dataset for Abstractive Summarization

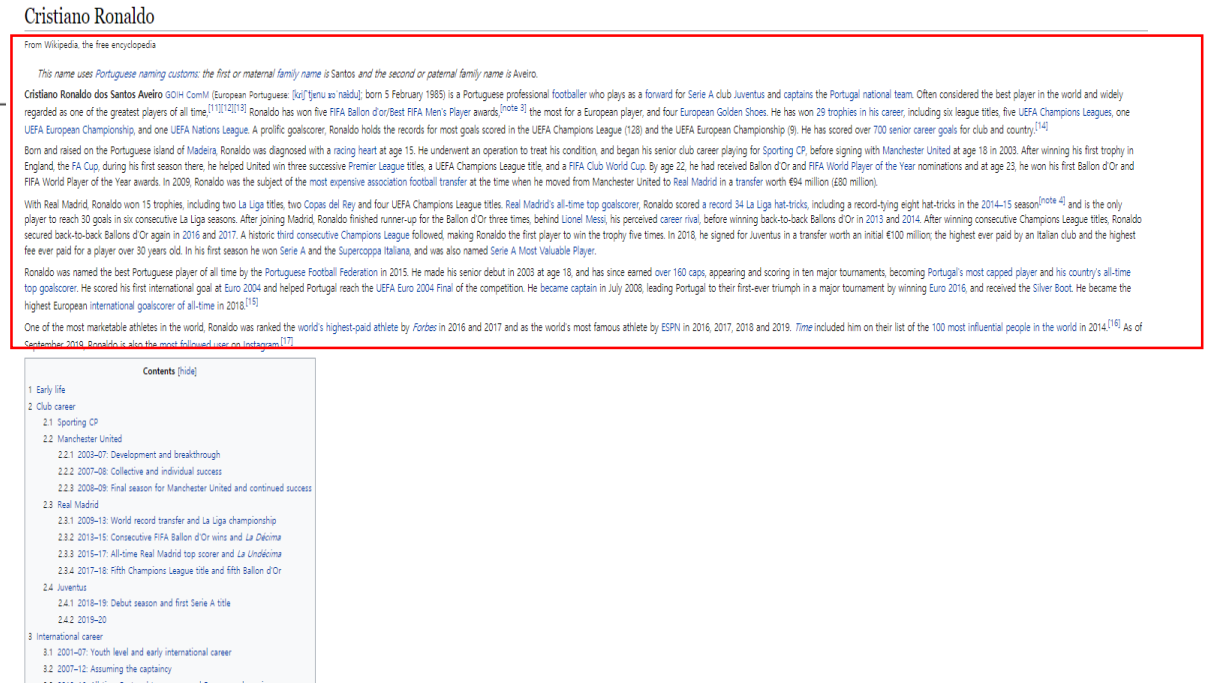
- English Gigaword Corpus
- Daily Mail and CNN

Table 1: Order of magnitude input/output sizes and unigram recall for summarization datasets.

Dataset	Input	Output	# examples	ROUGE-1 R
Gigaword (Graff & 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
WikiSum (ours)	10^2–10^6	10^1–10^3	10^6	59.2

Table 2: Percentiles for different aspects of WikiSum dataset. Size is in number of words.

Percentile	20	40	50	60	80	100
Lead Size	37	62	78	98	166	10,034
Num Citations	1	2	2	3	5	1,029
Citations Size	562	1,467	2,296	3,592	10,320	6,159,463
Num Search Results	10	20	26	31	46	2,095
Search Results Size	1,1691	33,989	49,222	68,681	135,533	5,355,671



Cristiano Ronaldo

From Wikipedia, the free encyclopedia

This name uses Portuguese naming customs: the first or maternal family name is Santos *and the second or paternal family name is* Aveiro.

Cristiano Ronaldo dos Santos Aveiro GOIH ComM (European Portuguese: [kɾiʃˈtjɐnu ʁɔˈnaldu]; born 5 February 1985) is a Portuguese professional footballer who plays as a forward for Serie A club Juventus and captains the Portugal national team. Often considered the best player in the world and widely regarded as one of the greatest players of all time,[11][12][13] Ronaldo has won five FIFA Ballon d'or/Best FIFA Men's Player awards,[note 3] the most for a European player, and four European Golden Shoes. He has won 29 trophies in his career, including six league titles, five UEFA Champions Leagues, one UEFA European Championship, and one UEFA Nations League. A prolific goalscorer, Ronaldo holds the records for most goals scored in the UEFA Champions League (128) and the UEFA European Championship (9). He has scored over 700 senior career goals for club and country.[14]

Born and raised on the Portuguese island of Madeira, Ronaldo was diagnosed with a racing heart at age 15. He underwent an operation to treat his condition, and began his senior club career playing for Sporting CP, before signing with Manchester United at age 18 in 2003. After winning his first trophy in England, the FA Cup, during his first season there, he helped United win three successive Premier League titles, a UEFA Champions League title, and a FIFA Club World Cup. By age 22, he had received Ballon d'Or and FIFA World Player of the Year nominations and at age 23, he won his first Ballon d'Or and FIFA World Player of the Year awards. In 2009, Ronaldo was the subject of the most expensive association football transfer at the time when he moved from Manchester United to Real Madrid in a transfer worth €94 million (£80 million).

With Real Madrid, Ronaldo won 15 trophies, including two La Liga titles, two Copas del Rey and four UEFA Champions League titles. Real Madrid's all-time top goalscorer, Ronaldo scored a record 34 La Liga hat-tricks, including a record-tying eight hat-tricks in the 2014–15 season[note 4] and is the only player to reach 30 goals in six consecutive La Liga seasons. After joining Madrid, Ronaldo finished runner-up for the Ballon d'Or three times, behind Lionel Messi, his perceived career rival, before winning back-to-back Ballons d'Or in 2013 and 2014. After winning consecutive Champions League titles, Ronaldo secured back-to-back Ballons d'Or again in 2016 and 2017. A historic third consecutive Champions League followed, making Ronaldo the first player to win the trophy five times. In 2018, he signed for Juventus in a transfer worth an initial €100 million; the highest ever paid by an Italian club and the highest fee ever paid for a player over 30 years old. In his first season he won Serie A and the Supercoppa Italiana, and was also named Serie A Most Valuable Player.

Ronaldo was named the best Portuguese player of all time by the Portuguese Football Federation in 2015. He made his senior debut in 2003 at age 18, and has since earned over 160 caps, appearing and scoring in ten major tournaments, becoming Portugal's most capped player and his country's all-time top goalscorer. He scored his first international goal at Euro 2004 and helped Portugal reach the UEFA Euro 2004 Final of the competition. He became captain in July 2008, leading Portugal to their first-ever triumph in a major tournament by winning Euro 2016, and received the Silver Boot. He became the highest European international goalscorer of all-time in 2018.[15]

One of the most marketable athletes in the world, Ronaldo was ranked the world's highest-paid athlete by *Forbes* in 2016 and 2017 and as the world's most famous athlete by ESPN in 2016, 2017, 2018 and 2019. *Time* included him on their list of the 100 most influential people in the world in 2014.[16] As of September 2019, Ronaldo is also the most followed user on Instagram.[17]

Contents [hide]

1 Early life
2 Club career
2.1 Sporting CP
2.2 Manchester United
2.2.1 2003–07: Development and breakthrough
2.2.2 2007–08: Collective and individual success
2.2.3 2008–09: Final season for Manchester United and continued success
2.3 Real Madrid
2.3.1 2009–13: World record transfer and La Liga championship
2.3.2 2013–15: Consecutive FIFA Ballon d'Or wins and *La Décima*
2.3.3 2015–17: All-time Real Madrid top scorer and *La Undécima*
2.3.4 2017–18: Fifth Champions League title and fifth Ballon d'Or
2.4 Juventus
2.4.1 2018–19: Debut season and first Serie A title
2.4.2 2019–20
3 International career
3.1 2001–07: Youth level and early international career
3.2 2007–12: Assuming the captaincy

WikiSum

For given a_i (Wikipedia Ariticle)

1. Cited Sources : C_i
2. Web Search Results : S_i (Reference가 부족한 경우)

Table 2: Percentiles for different aspects of WikiSum dataset. Size is in number of words.

Percentile	20	40	50	60	80	100
Lead Size	37	62	78	98	166	10,034
Num Citations	1	2	2	3	5	1,029
Citations Size	562	1,467	2,296	3,592	10,320	6,159,463
Num Search Results	10	20	26	31	46	2,095
Search Results Size	1,1691	33,989	49,222	68,681	135,533	5,355,671

Training Method

1. Extractive Summarization(Due to HW problem)
   - -First L tokens
   - -TF-IDF
   - -SumBasic
   - -TextRank
   - -Cheating
2. Abstractive Summarization
   - -Transofmer Decoder

Extractive Method

1. First L tokens
2. TF-IDF : Query(Title)과 가장 유사한 Paragraph 추출
3. TextRank : Similar as PageRank

- paragraph : node
- edge : similarity based on word overlap

4. SumBasic
5. Cheating : Recall of bigram of ground truth

$$d(p_j^i, a_i) = \frac{bigrams(p_j^i) \cap bigrams(a_i)}{bigrams(a_i)}$$

Sumbasic

Step 1 Compute the probability distribution over the words w_i appearing in the input, $p(w_i)$ for every i; $p(w_i) = \frac{n}{N}$, where n is the number of times the word appeared in the input, and N is the total number of content word tokens in the input.

Step 2 For each sentence S_j in the input, assign a weight equal to the average probability of the words in the sentence, i.e.

$$weight(S_j) = \sum_{w_i \in S_j} \frac{p(w_i)}{|\{w_i | w_i \in S_j\}|}$$

Step 3 Pick the best scoring sentence that contains the highest probability word.

Step 4 For each word w_i in the sentence chosen at step 3, update their probability

$$p_{new}(w_i) = p_{old}(w_i) \cdot p_{old}(w_i)$$

Step 5 If the desired summary length has not been reached, go back to Step 2.

- Step 4는 Summary에 중복된 단어가 등장할 확률을 낮춰준다
- Low Probability의 word도 뽑힐 확률이 생김

Abstractive Stage

- Given Title and Ordered paragraph

$$text_i = T(a_i) \| \{p^i_{R_i(j)}\}$$

$$tokenize(text_i) = x_i = (x_i^1, x_i^2, ..., x_i^{n_i})$$

- Truncating

$$m_i^L = (x_i^1, ...x_i^{min(L,n_i)})$$

- Learning

$$a_i = W(m_i^L)$$

Models

- Seq2Seq with attention
- Transformer ED
- Transformer Decoder Only

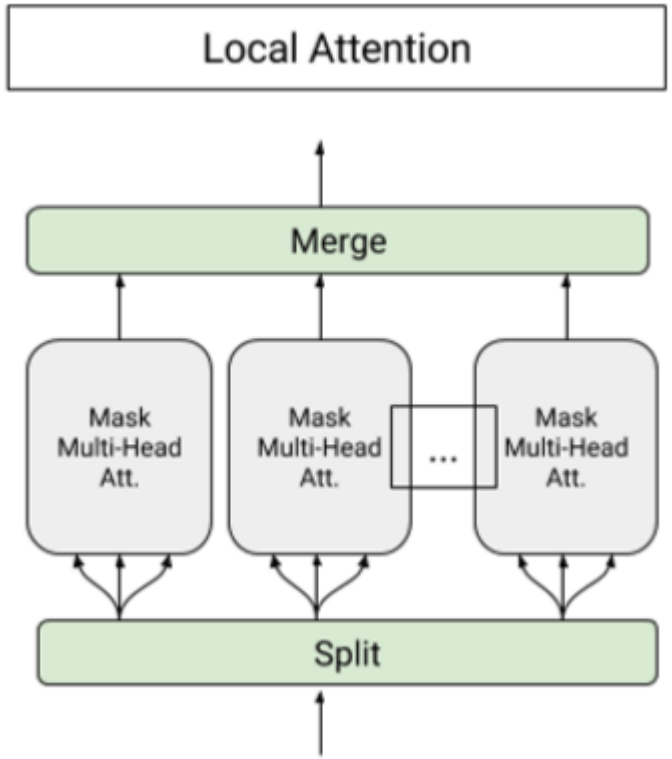
$$(m^1, ..., m^n) \mapsto (y^1, ..., y^\eta)$$

$$(w^1, ..., w^{n+\eta+1}) = (m^1, ..., m^n, \delta, y^1, ..., y^\eta)$$

Local Attention

- Local Attention

: subsequence로 나누고 subsequence내에서 multihead attention



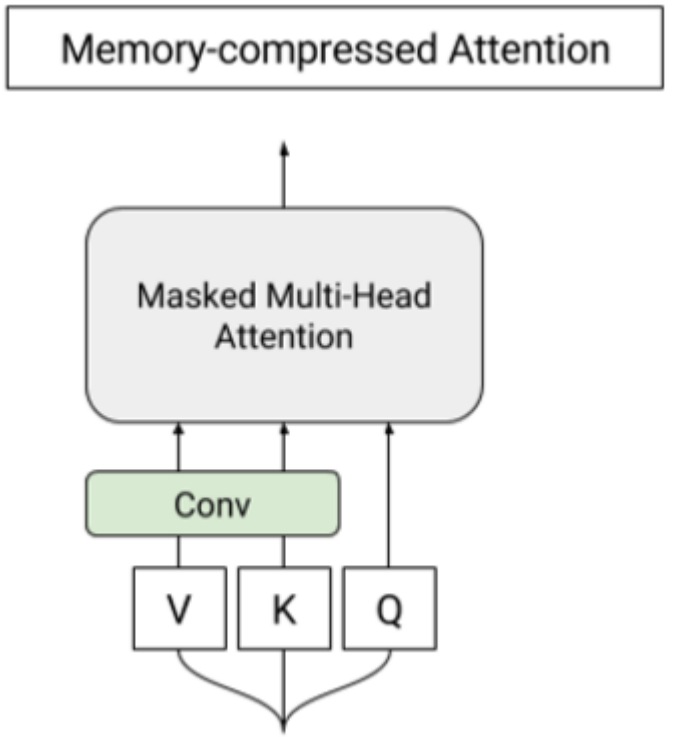
Memory Compressed Attention

- Reduce number of key and values

(Number fo queries는 그대로)

- Global Information 반영
- Final Architecture : 5 layers(LMLML)

In our experiments we use convolution kernels of size 3 with stride 3



Experiments

- Perplexity
- ROUGE-F1
- ROUGE-Recall : Long summarization
- ROUGE-Precision : Short summarization

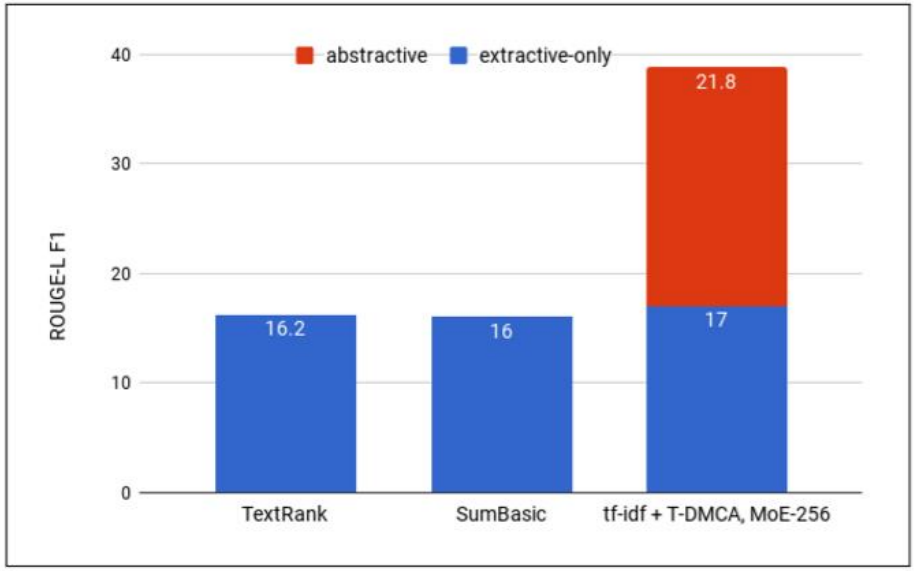


Figure 2: ROUGE-L F1 for various extractive methods. The abstractive model contribution is shown for the best combined *tf-idf*-T-DMCA model.

Table 4: Performance of best models of each model architecture using the combined corpus and tf-idf extractor.

Model	Test perplexity	ROUGE-L
seq2seq-attention, $L = 500$	5.04952	12.7
Transformer-ED, $L = 500$	2.46645	34.2
Transformer-D, $L = 4000$	2.22216	33.6
Transformer-DMCA, no MoE-layer, $L = 11000$	2.05159	36.2
Transformer-DMCA, MoE-128, $L = 11000$	1.92871	37.9
Transformer-DMCA, MoE-256, $L = 7500$	1.90325	38.8

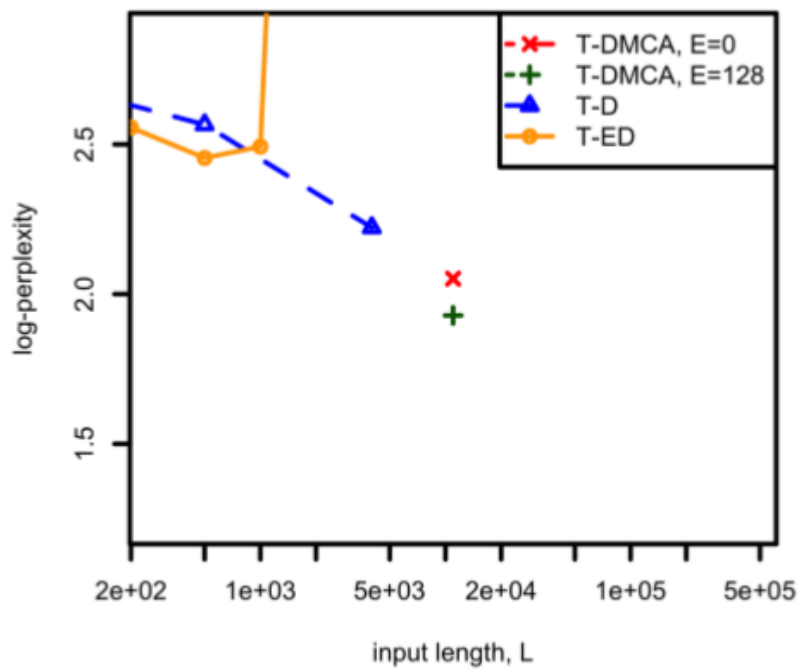


Figure 3: Shows perplexity versus L for tf-idf extraction on combined corpus for different model architectures. For T-DMCA, E denotes the size of the mixture-of-experts layer.

Transformer-encoder-decoder, *L*=100 (log-perplexity: 2.63)
dewey & leboeuf llp (dewey & leboeuf llp) is an american law firm headquartered in new york city . dewey & leboeuf is one of the largest law firms in the united states . dewey & leboeuf has offices in new york city , los angeles , washington , d.c. , washington , d.c. , and washington , d.c.

Transformer decoder, *L*=500 (log-perplexity: 2.60)
dewey & leboeuf llp is an international law firm headquartered in new york city . dewey was formed in october 2007 through the combination of dewey ballantine llp and leboeuf , lamb , greene , & macrae llp .

Transformer-DMAC, L=7000, 256 experts (log-perplexity: 1.90)
dewey & leboeuf llp is an international law firm headquartered in new york city . it was formed in october 2007 through the combination of dewey ballantine llp and leboeuf , lamb , greene & macrae llp . at its height , approximately 1,300 partners and employees worked in dewey 's manhattan office , and nearly 3,000 partners and employees worked for the firm worldwide . in may 2012 , dewey collapsed , resulting in the largest law firm bankruptcy

Wikipedia (ground truth)
dewey & leboeuf llp was a global law firm , headquartered in new york city , that is now in bankruptcy . the firm 's leaders have been indicted for fraud for their role in allegedly cooking the company 's books to obtain loans while hiding the firm 's financial plight . the firm was formed in 2007 through the merger of dewey ballantine and leboeuf , lamb , greene & macrae . dewey & leboeuf was known for its corporate , insurance , litigation , tax and restructuring practices . at the time of the bankruptcy filing , it employed over 1,000 lawyers in 26 offices around the world . in 2012 , the firm 's financial difficulties and indebtedness became public . in the same period , many partners departed , and the manhattan district attorney 's office began to investigate alleged false statements by firm chairman steven davis . as a result of these difficulties , dewey & leboeuf 's offices began to enter administration in may 2012 . the firm filed for bankruptcy in new york on may 28 , 2012 . on march 6 , 2014 , the former chairman , chief financial officer and the executive director of dewey & leboeuf were indicted on charges of grand larceny by the manhattan district attorney .

Figure 4: Shows predictions for the same example from different models. Example model input can be found in the Appendix A.4

Generating Full Wikipedia

One is trained to use L = 6000 reference tokens to predict at most 2192 article tokens(longer examples are ignored)

and another is conditioned only on the title and generates articles up to 4000 tokens long.

A.1 EXAMPLES OF FULL WIKIPEDIA GENERATED SAMPLES

OUTPUT:
= = Wings Over Kansas

= =wings over kansas is the best aviation history website i have encountered on the world wide web . it is informative , entertaining , provides ever changing content , and is populated with the true voices of the mainstream aviation community . there is no better place to see where aviation has been and where it is going . centered in the midst of the greatest producer of aircraft in the world , wings over kansas reflects that aviation community to the rest of the world .

= = Wings Over Kansas History

= =wings over kansas was established with the mission of becoming the number one online kansas aviation resource showcasing the pioneers , educators , newsmakers , manufacturers , pilots and craftsmen , who have made kansas the world center for aviation production . wings over kansas was established with the mission of becoming the number one online kansas aviation resource showcasing the pioneers , educators , newsmakers , manufacturers , pilots and craftsmen , who have made kansas the world center for aviation production . wings over kansas has been recognized by mcgraw - hill as one of the top 500 best aviation web sites with visitors from over 225 countries . wings over kansas offers a unique perspective on the role of wichita and kansas in the history and development of international aviation . the featured menu offers over 1,500 pages on aviation news , history , education , photos , videos , careers , pioneers , quizzes and learn - to - fly categories . in addition , the special subjects section offers further aviation content pages to visit .

= = Wings Over Kansas Features

= =wings over kansas offers a unique perspective on the role of wichita and kansas in the history and development of international aviation . the featured menu offers over 1,500 pages on aviation news , history , education , photos , videos , careers , pioneers , quizzes and learn - to - fly categories . in addition , the special subjects section offers further aviation content pages to visit .

TARGET:
= = Wings Over Kansas

= =wings over kansas.com is an aviation website founded in 1998 by carl chance owned by chance communications , inc. to provide information and entertainment to aviation enthusiasts and professionals worldwide . the web site is based in wichita , kansas , known as the " air capital of the world " due to the many aircraft manufacturers located there . in 2003 , the site was upgraded to a data - based web site to better serve the needs of its members . " wings over kansas " has grown steadily and as of 2009 draws over a quarter of a million visitors yearly from over 125 countries .

= = Wings Over Kansas History

= =wings over kansas.com was created in 1998 by wichita native carl chance , a broadcast professional and producer for the wingspan air & space channel . in his more than thirty years of experience , chance developed many relationships in the aviation community that have directly benefited the web site . he is a charter member and past trustee on the kansas aviation museum board of directors and a former member of the kansas aviation council . from 1998 to 2003 , the site underwent a number of modifications to improve its value and navigation .

= = Wings Over Kansas History 2003 Redesign

= =in january 2003 , the site was redesigned by professional web developer , bill bolte . the new design included a data - based implementation to better serve the needs of the members including aviation professionals , educators , historians , and enthusiasts .

= = Wings Over Kansas Overview

= =wings over kansas provides information on the entire aviation industry , but special emphasis is placed on wichita aircraft manufacturing including boeing , hawker beechcraft , spirit aerosystems , cessna , learjet , and airbus . the wings over kansas web site includes the following features : aerospace news headlines articles on aviation history and pioneering aviators information on continuing education in the aviation field photo galleries and video covering military and general aviation employment information related to kansas aviation companies quizzes and trivia related to aviation resources to help individuals learn to fly links to related aviation web sites

= = Wings Over Kansas Contributing editors

= =wings over kansas receives support from a diverse group of contributing editors including : walter j. boyne – aviation author and historian ; former director of the smithsonian national air and space museum lionel

Figure 6: An example decoded from a T-DMCA model trained to produce an entire Wikipedia article, conditioned on 8192 reference document tokens.