News2Images: Automatically Summarizing News Articles into Image-Based Contents via Deep Learning

20 September 2015

The 3rd International Workshop on News Recommendation and Analytics (INRA 2015)

Authors: Jung Woo Ha, Dongyeop Kang, Hyuna Pyo, and Jeonghee Kim

NAVER LABS

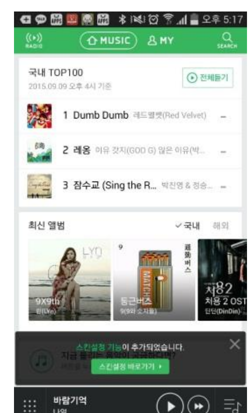
E-mail: jungwoo.ha@navercorp.com

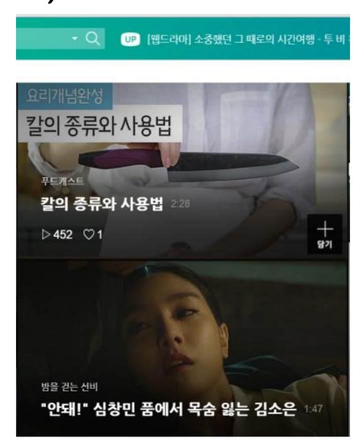
NAVER LABS

Contents

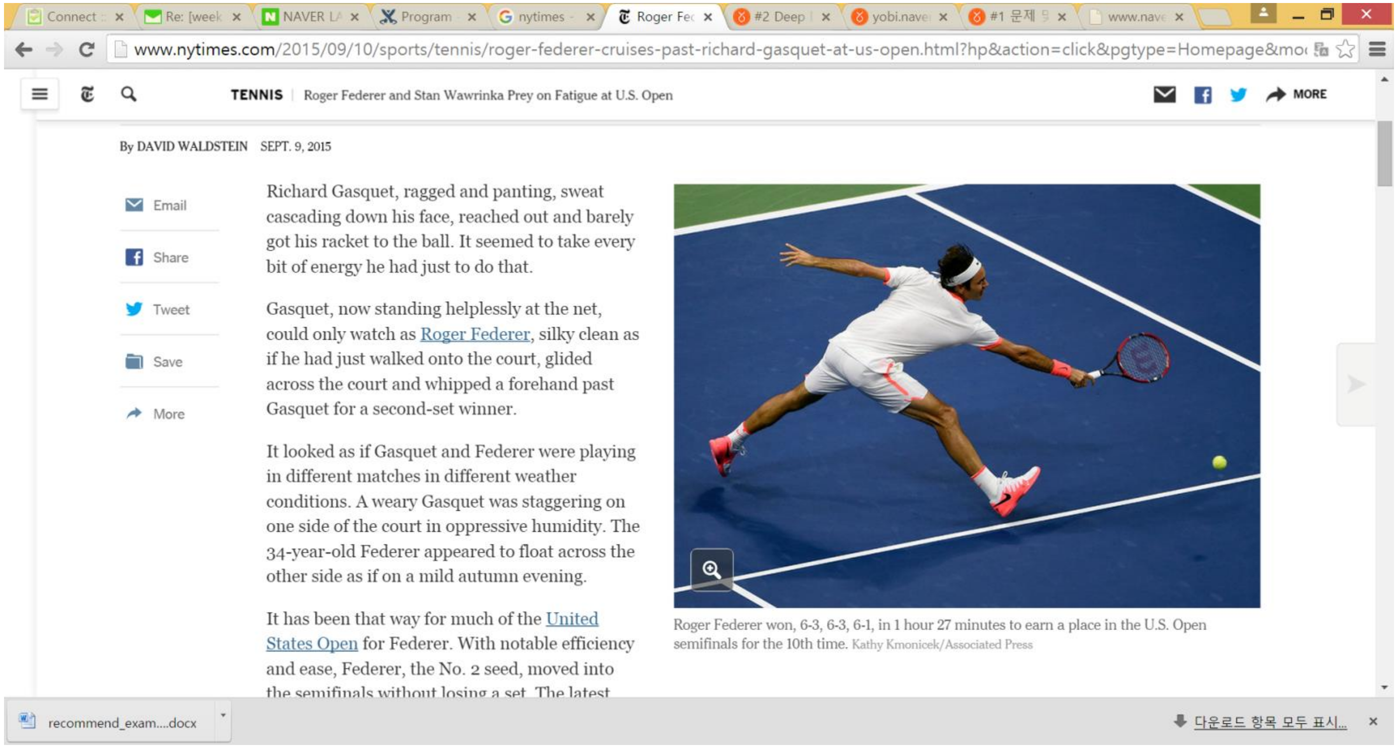
NAVER

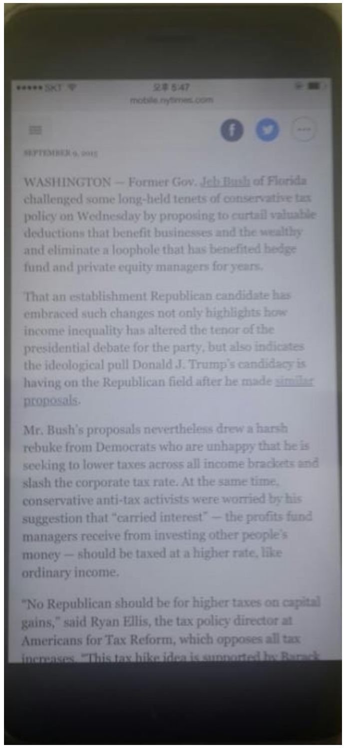
- No. 1 internet company of Korea (www.naver.com)
  - Web portal, global messenger, music, podcast, video, collective intelligence, search advertisement, app store
  - Line messenger MAU: globally 250 million
  - Daily mobile page views: 1,632 million
  - Daily search page views: 504 million
- NAVER LABS: R&D center for advanced technologies
  - Machine learning, recommendation, speech recognition, machine translation, multimedia, web browser, IOT, HPC infra, etc.





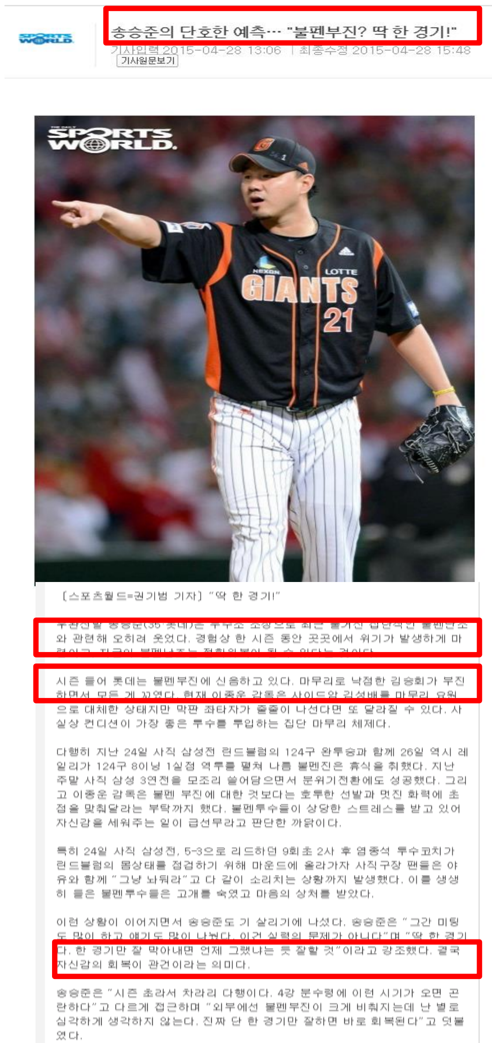
Why Post-like News?





Post-like News Summarization

[Online news article]



송승준의 단호한 예측… "불펜부진? 딱 한 경기!"

기사입력 2015-04-28 13:06 | 최종수정 2015-04-28 15:48

기사원문보기

[스포츠월드=권기범 기자] "딱 한 경기!"

와 관련해 오히려 웃었다. 경험상 한 시즌 동안 곳곳에서 위기가 발생하게 마

시즌 들어 롯데는 불펜부진에 신음하고 있다. 마무리로 낙점한 김승회가 부진하면서 모든 게 꼬였다. 현재 이종운 감독은 사이드암 김성배를 마무리 요원으로 대체한 상태지만 막판 좌타자가 줄줄이 나선다면 또 달라질 수 있다. 사실상 컨디션이 가장 좋은 투수를 투입하는 집단 마무리 체제다.

다행히 지난 24일 사직 삼성전 린드블럼의 124구 완투승과 함께 26일 역시 레일리가 124구 8이닝 1실점 역투를 펼쳐 나름 불펜진은 휴식을 취했다. 지난 주말 사직 삼성 3연전을 모조리 쓸어담으면서 분위기전환에도 성공했다. 그리고 이종운 감독은 불펜 부진에 대한 것보다는 호투한 선발과 멋진 화력에 초점을 맞춰달라는 부탁까지 했다. 불펜투수들이 상당한 스트레스를 받고 있어 자신감을 세워주는 일이 급선무라고 판단한 까닭이다.

특히 24일 사직 삼성전, 5-3으로 리드하던 9회초 2사 후 염종석 투수코치가 린드블럼의 몸상태를 점검하기 위해 마운드에 올라가자 사직구장 팬들은 야유와 함께 "그냥 놔둬라"고 다 같이 소리치는 상황까지 발생했다. 이를 생생히 들은 불펜투수들은 고개를 숙였고 마음의 상처를 받았다.

이런 상황이 이어지면서 송승준도 기 살리기에 나섰다. 송승준은 "그간 미팅도 많이 하고 얘기도 많이 나눴다. 이건 실력의 문제가 아니다"며 "딱 한 경기다. 한 경기만 잘 막아내면 언제 그랬냐는 듯 잘할 것"이라고 강조했다. 결국 자신감의 회복이 관건이라는 의미다.

송승준은 "시즌 초라서 차라리 다행이다. 4강 분수령에 이런 시기가 오면 곤란하다"고 다르게 접근하며 "외부에선 불펜부진이 크게 비춰지는데 난 별로 심각하게 생각하지 않는다. 진짜 단 한 경기만 잘하면 바로 회복된다"고 덧붙였다.

[Summarized content]

Post-like News Summarization

- Task definition
  - Summarizing a document into multiple image-based contents
  - Pikicast: https://www.pikicast.com/

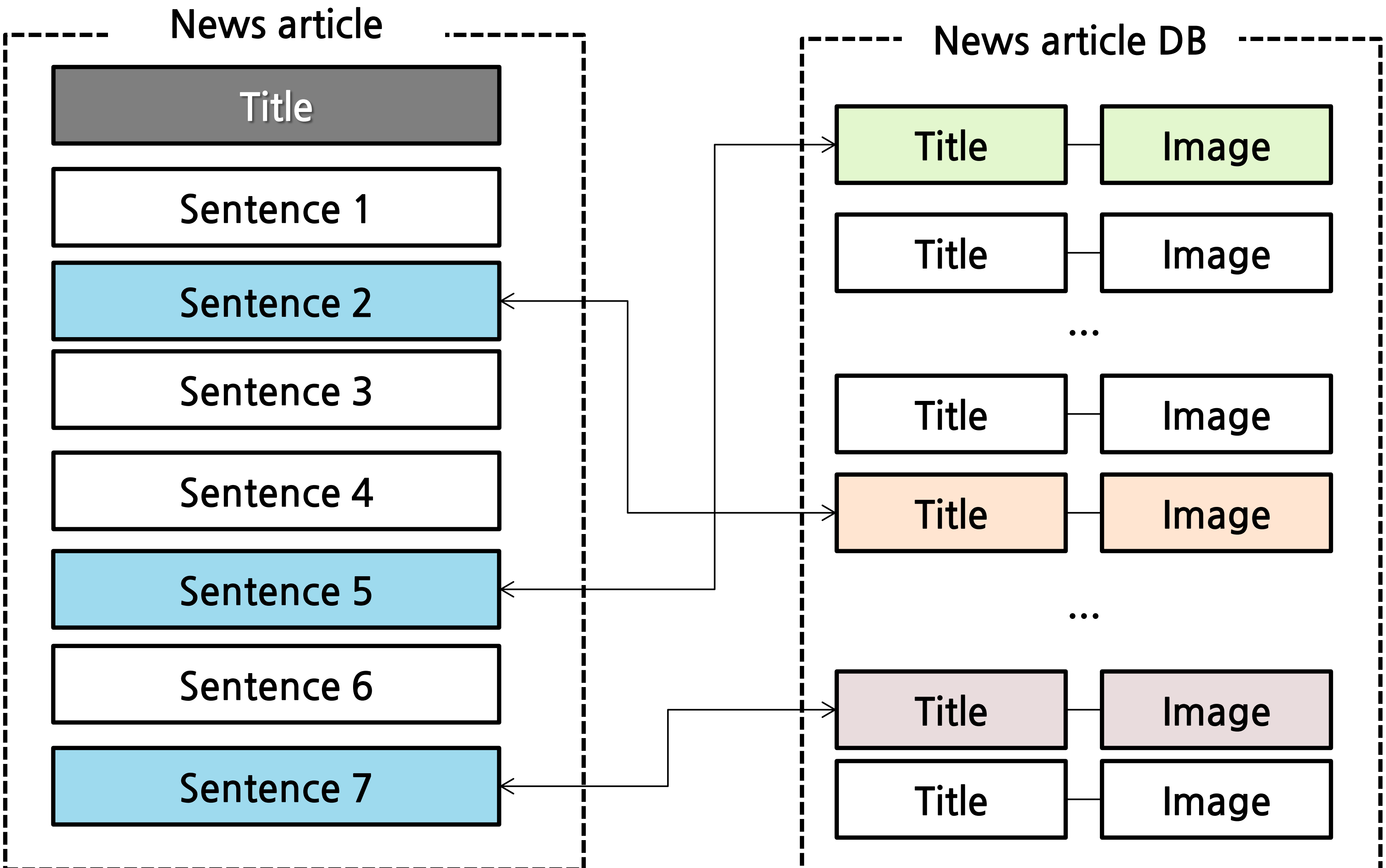
- Significance
  - Overcoming the display size of mobile devices
  - Summarization into image-based contents instead of texts
    - Easy to see the news
    - Enhancing users' interests
  - Multimodal documents such as blogs as well as news articles
  - Applied to visual-linguistic cross-modal transformation

Post-like News Summarization

- Subtasks
  - Document summarization
  - Text-to-image retrieval
  - Image-based content generation

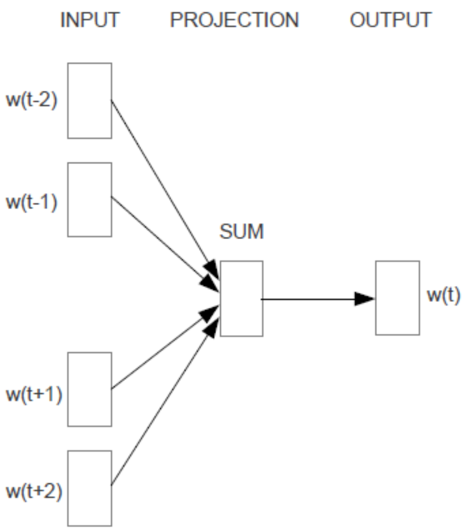
- Key technologies
  - Document data embedding based on deep learning
  - Document summarization considering two factors: similarity and diversity
  - Image feature generation using convolutional neural networks (CNNs)
  - Common semantic embedding-based text-image association

Post-like News Summarization

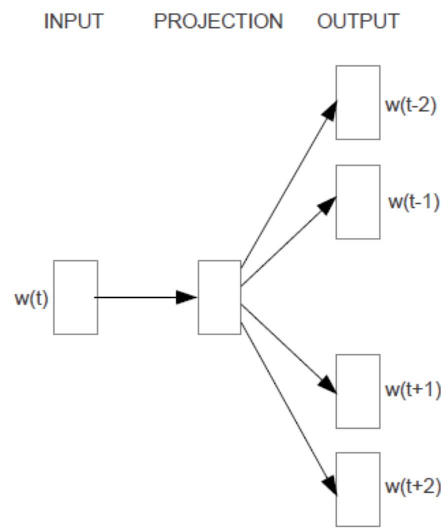


News2Images

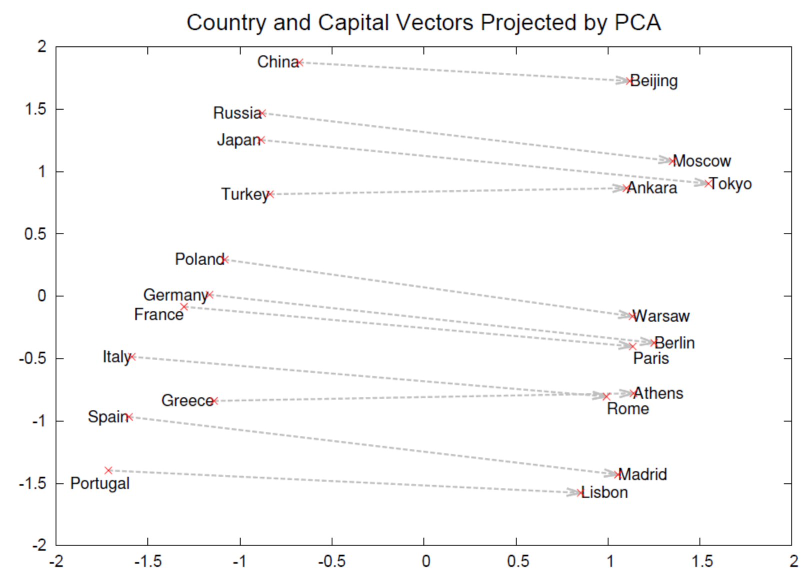
- Deep learning-based text feature representation
  - Word2Vec (Mikolov et al. 2013)
  - Words (paragraphs, documents) → real-valued vectors
  - Neural network model for distributed representation



CBOW



Skip-gram



News2Images

- Document summarization
  - Key sentence extraction
  - A document → a sentence set
  - Selecting *k* sentences covering the document contents
  - Criteria: similarity and diversity
    - $f(S_k, S)$: similarity with the title
    - $g(S_k, S)$: diverse words for semantic coverage

$$S_k^* = \underset{S_k \subset S}{\arg\max} \left\{ \alpha \cdot f\left(S_k, S\right) + \left(1-\alpha\right) \cdot g\left(S_k, S\right) \right\} \qquad f\left(S_k, S\right) = \sum\nolimits_{\mathbf{s} \in S_k} f\left(\mathbf{s}, S\right)$$

$$= \underset{S_k \subset S}{\arg\max} \left\{ \alpha \cdot f\left(S_k, t\right) + \left(1-\alpha\right) \cdot g\left(S_k, S\right) \right\} \qquad g\left(S_k, S\right) = \sum\nolimits_{\mathbf{s} \in S_k} g\left(\mathbf{s}, S\right)$$

  - α : a constant for moderating the similarity and the diversity
  - S_k: the set of *k* selected sentences as the summarization of *S*

- Two approaches
  - Baseline: TF / IDF + Word occurrence vector of sentences
  - Sentence embedding

News2Images

- TF / IDF
  - Similarity: cosine similarity between two vectors

$$f(\mathbf{s}, S) = f(\mathbf{s}, \mathbf{t}) = \cos\mathrm{sim}(\mathbf{s}, \mathbf{t}) = \frac{\mathbf{s} \cdot \mathbf{t}}{\| \mathbf{s} \| \| \mathbf{t} \|}$$

    - **s** and **t**: word occurrence vectors of a sentence and the title of a given article including **s**
  - Diversity: prefer sentences including diverse words

$$g(\mathbf{s}, S) = \frac{1}{|S| \cdot \prod_{\mathbf{s} \neq \mathbf{s}', \mathbf{s}' \in S} \{\cos\mathrm{sim}(\mathbf{s}, \mathbf{s}')\}}$$

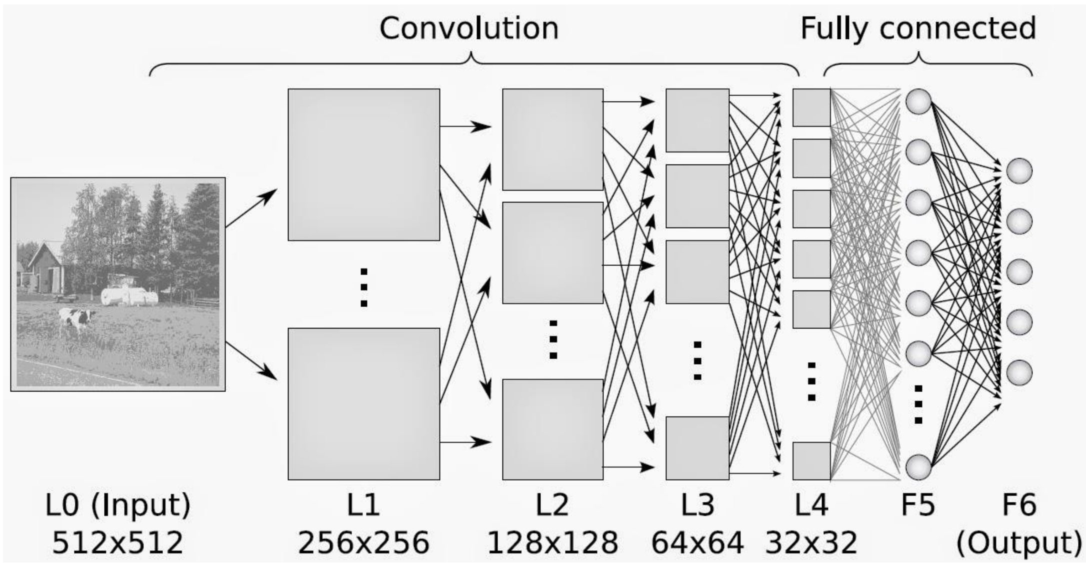
- Sentence embedding
  - |**s**| = 100 and average pooling on all the word vectors in **s**
  - Similarity: cosine similarity between two embedding vectors
  - Diversity: an article → clusters → representative sentences of each cluster

$$g(\mathbf{s}, S) = g(\mathbf{s}, C^j) = \cos\mathrm{sim}(\mathbf{s}, \mathbf{c}^j) = \frac{\mathbf{s} \cdot \mathbf{c}^j}{\| \mathbf{s} \| \| \mathbf{c}^j \|}$$

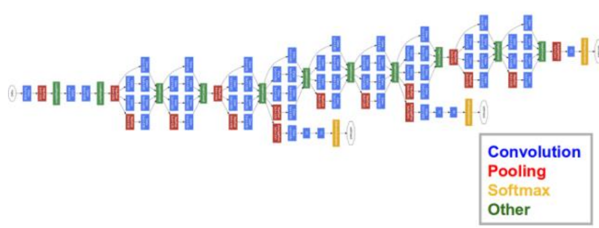
    - C^j and c^j: the *j*-th cluster and its centroid, $\mathrm{s} \in C^j$

News2Images

- How to generate image features?
  - Deep convolutional neural networks (Krizhevsky et al. 2012)



News images



GoogleNet in Caffe

{0.231, …, 1.234}

Real-valued vectors

News2Images

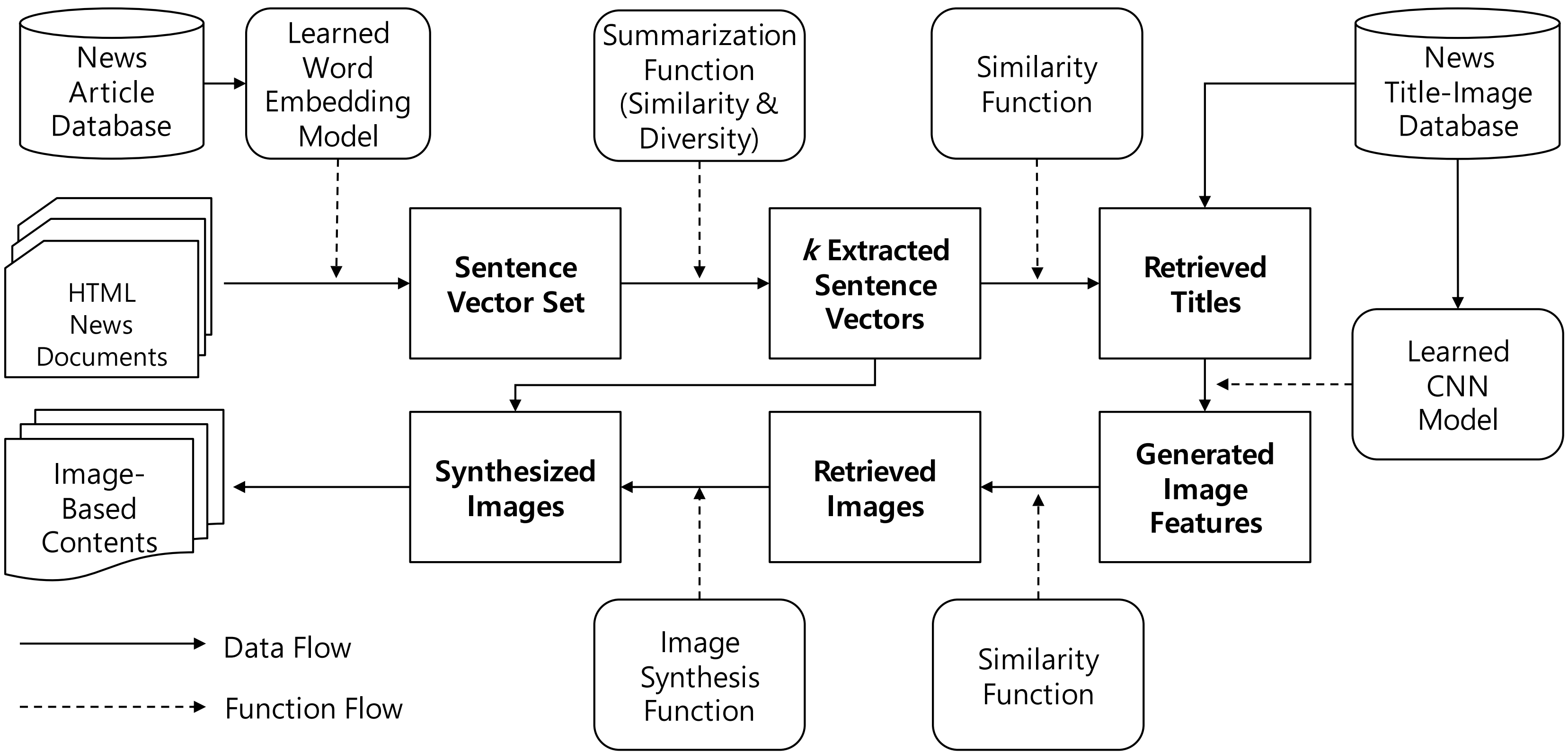
- Text-to-image retrieval
  - Image feature generation
    - Modified GoogleNet implented in Caffe (Jia et al. 2014)
    - Fully connected 1024 dim features
    - For supervised learning, each image is labeled with the name included in its title
  - Algorithms

$$\mathbf{v}^* = \arg\max_{\mathbf{v}\in V} \left\{ f\left(\hat{\mathbf{s}}, \mathbf{t}(\mathbf{v})\right) \right\} = \arg\max_{\mathbf{v}\in V} \left\{ \frac{\hat{\mathbf{s}} \cdot \mathbf{t}(\mathbf{v})}{|\hat{\mathbf{s}}||\mathbf{t}(\mathbf{v})|} \right\}$$

    - Method 1: Text → most similar text → image
    - Method 2: Text → most similar *N* texts → new image features → most similar image
      - A new image feature is generated by averaging features of top N images whose title is similar to the extracted sentence
      - Retrieve the image most similar to the generated feature
  - Baseline: the cosine similarity between word occurrence vectors of the news title of an candidate image and the extracted sentences

News2Images

- Overall flow of News2Images

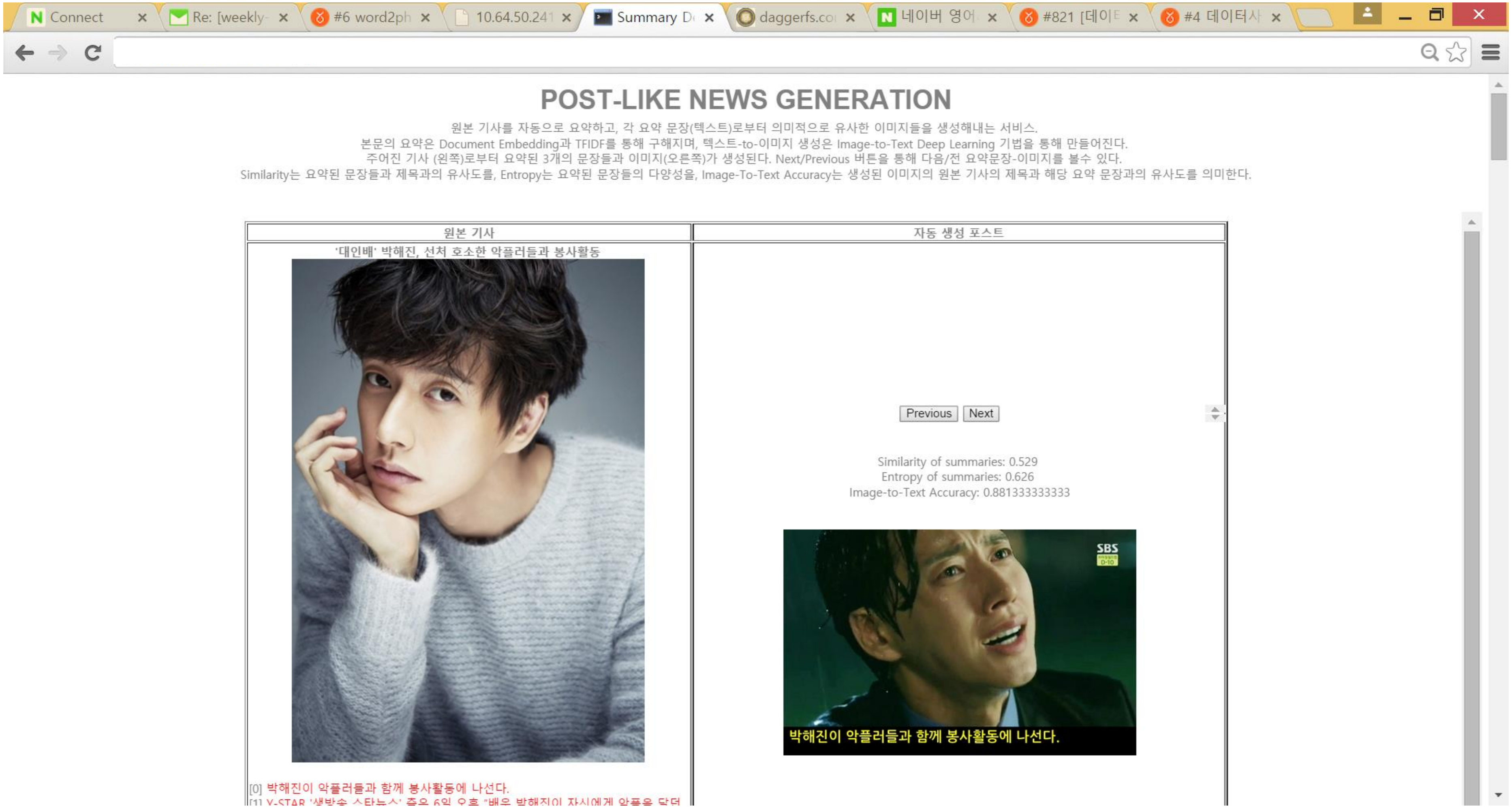


Experimental Results

- Data description
  - Training data
    - Word embedding: 1.1 million news articles on sports / entertainment in 2014 of NAVER
    - CNNs for Image features: 0.23 million photo images attached in NAVER news articles including 100 movie/sports starts
  - Evaluation data
    - Image features : 60,000 photo images
    - News summarization: 7,000 news articles on 100 movie/sports starts
- Parameter setup
  - Word embedding
    - Vector size: 100
  - Summarization: three sentences for each news article
  - Image features: GoogleNet 650,000 iteration
  - Image classification accuracy (100 classes)
    - Top 1: 54.8% / Top 5: 75.8%

Experimental Results

- Demo page



Experimental Results

- Classification
  - Correct case: An image includes the person referred in a summarized sentence
- Comparison of two T2I methods

Classification	Base-line (With title)	Text matching (M1)		Image averaging (M2)	
		With title	W/O title	With title	W/O title
Correct	14910	18908	13896	18791	13860
Accuracy(%)	73.7	**93.5**	68.7	92.9	68.5

  - With title: the title of the summarized news is given with the summarized sentence together
  - M2 > M1: feature averaging → vanishing of the unique properties of each image

Experimental Results

- Weighted to person name
  - Noun of a person name is weighted when pooling text vectors

Classification	PS weight = 1.0		PS weight = 10.0	
	With title	W/O title	With title	W/O title
Correct	18908	13896	19191	14065
Accuracy(%)	0.934929	0.687104	**0.948922**	0.695461

- Size of a word window in vector pooling

Classification	Window size = 3		Window size = 1	
	M1	M2	M1	M2
Correct	18743	18557	19191	18791
Accuracy(%)	0.92677	0.917573	**0.934929**	0.929144

Experimental Results

- Examples of summarized sentences and retrieved images

Sentences	News2Images	Baseline
Park, the home run leader of KBO, hit the 34th home run in this season.		
Son of Leverkusen played as a starter forward in this game for 60 minutes until substituted with Yurchenko		
Today, Ryu pitched 7 innings, allowed two runs and 9 hits, and got 7 kills against the Chicago Cubs at the home game, and thus ERA becomes 3.39.		
Lee, Hyori is practicing yoga with a grave look in the released photo.		
Chu, Soohyun showed her bodyline at the swimming pool scene in the 18th episode of the drama.		

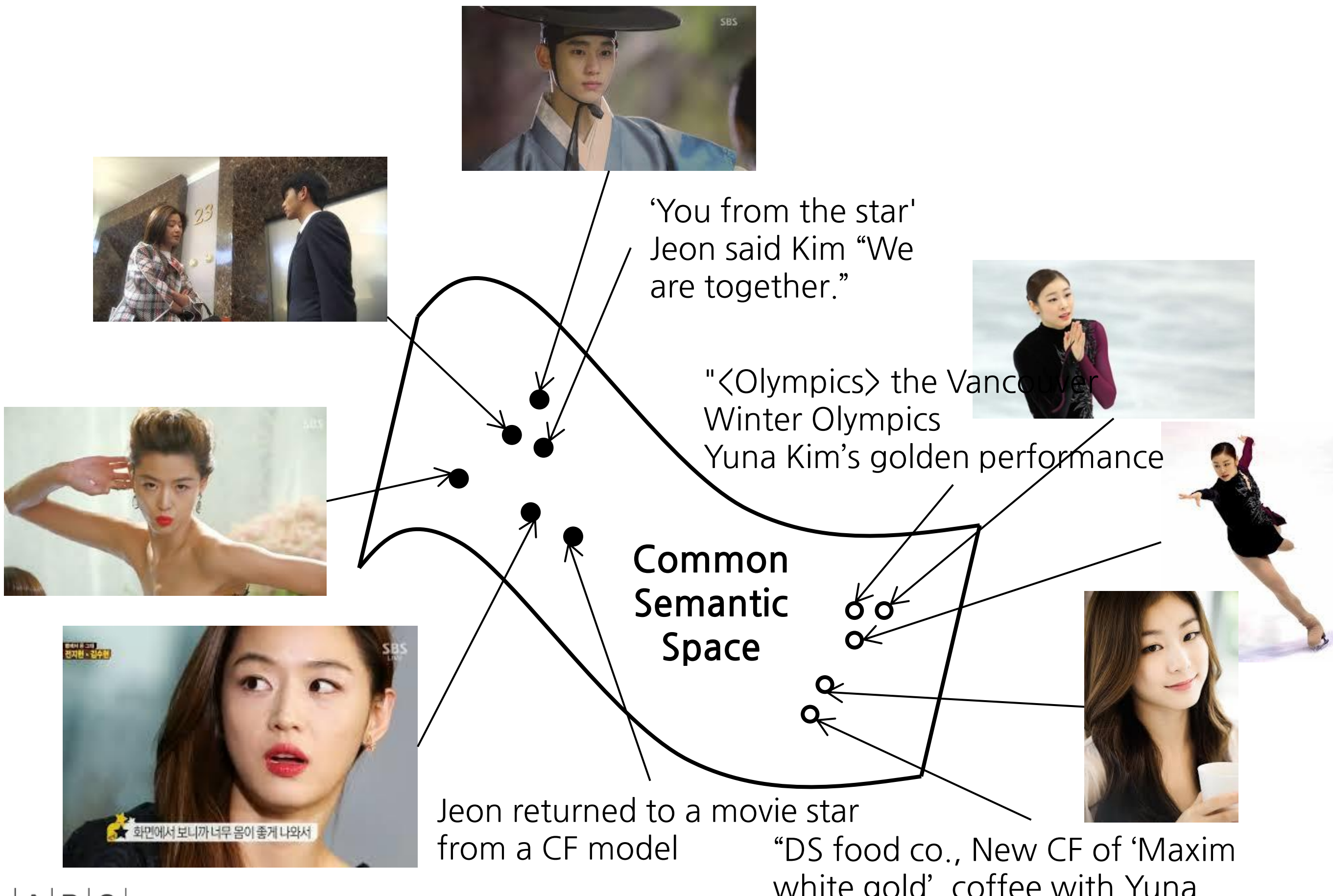
Discussion

- How to integrate News2Images with recommendation system
  - Preferred keywords → vector pooling and image retrieval
  - News document embedding → content-based recommendation for item cold start
  - Document-Image embedding → CF latent features: Hybrid recommendation (Van den Oord et al. 2013)

- Future work
  - An end-to-end model for text-image embedding
  - Thin implementation for mobile services
  - More articles and diverse subjects (politics, economy, society, etc.)
  - Integrating News2images with recommendation and personalization

Discussion and Future work

- End-to-end model for learning common semantic space from news-image data



References

1. Irsoy, O. and Cardie C., Deep recursive neural networks for compositionality in language. In *Advances in Neural Information Processing Systems* 2014. 2096-2104.
2. Jia, Y. et al. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia* 2014. 675-678.
3. Krizhevsky, A., Sutskever, I., and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 2012. 1097-1105.
4. LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature*. 521, 7553. 436-444.
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* 2013. 3111-3119.
6. Socher, R., Lin, C. C.-Y., Ng, A., and Manning, C. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning* (*ICML*-11). 129-136.
7. Van den Oord, A., Dieleman, S., and Schrauwen, B. 2013. Deep content-based music recommendation, In *Advances in Neural Information Processing Systems* 2013. 2643-2651.

Q&A