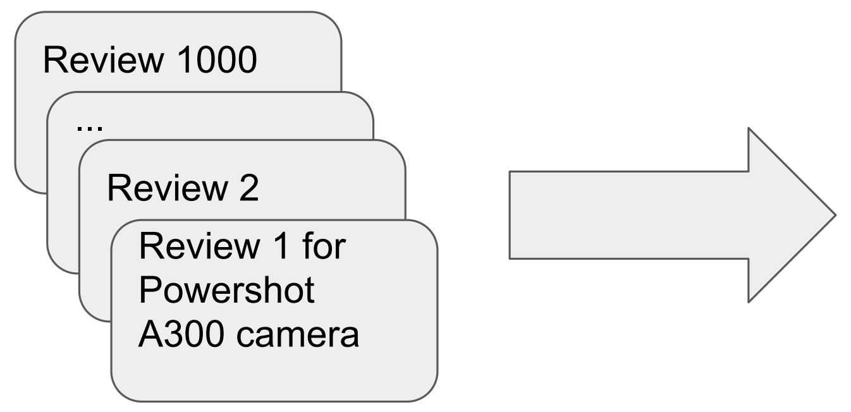
Recognizing themes in Amazon reviews through Unsupervised Multi-Document Summarization

Hanoz Bhathena, Jeff Chen, William Locke

{hvb2108, jc1,wlocke}@stanford.edu - December 2018

Motivation

- Number of reviews for products on Amazon is overwhelming
- Ideally, the shopper wants to have a summary of the themes
- Unsupervised learning to summarize themes across all reviews with a count of how many users mentioned every theme



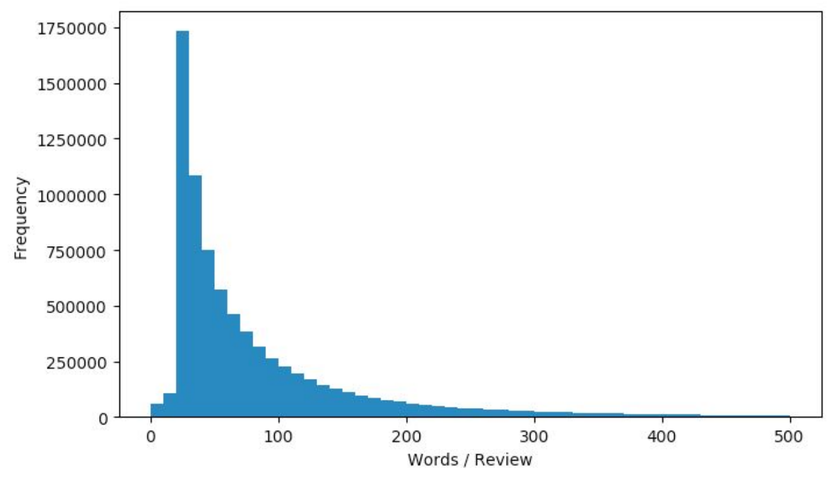
Input

Theme	Mentioned by
The camera has a good zoom on it, and is very easy to use.	37 people
It goes through batteries in about 10 minutes of photography.	22 people
Excellent value for money.	15 people

Output

Datasets

- Amazon reviews dataset [1]
- Electronics category with 7.8 Million reviews for 476k products
- Each review has a review text and a 0 to 5 score



Frequency of reviews vs word count. 90% of the reviews have less than 200 words.

Review	Score
Picture quality is very good. Eats batteries, but you need to buy rechargeables anyway, for any digicam, so it's not a big problem in my opinion. My camera says sometimes that batteries are depleted when they are not, when I turn it off and again on it works again.	5.0

Example review for a Powershot A300 Camera. 75 reviews total for this product.

Metrics & Baseline

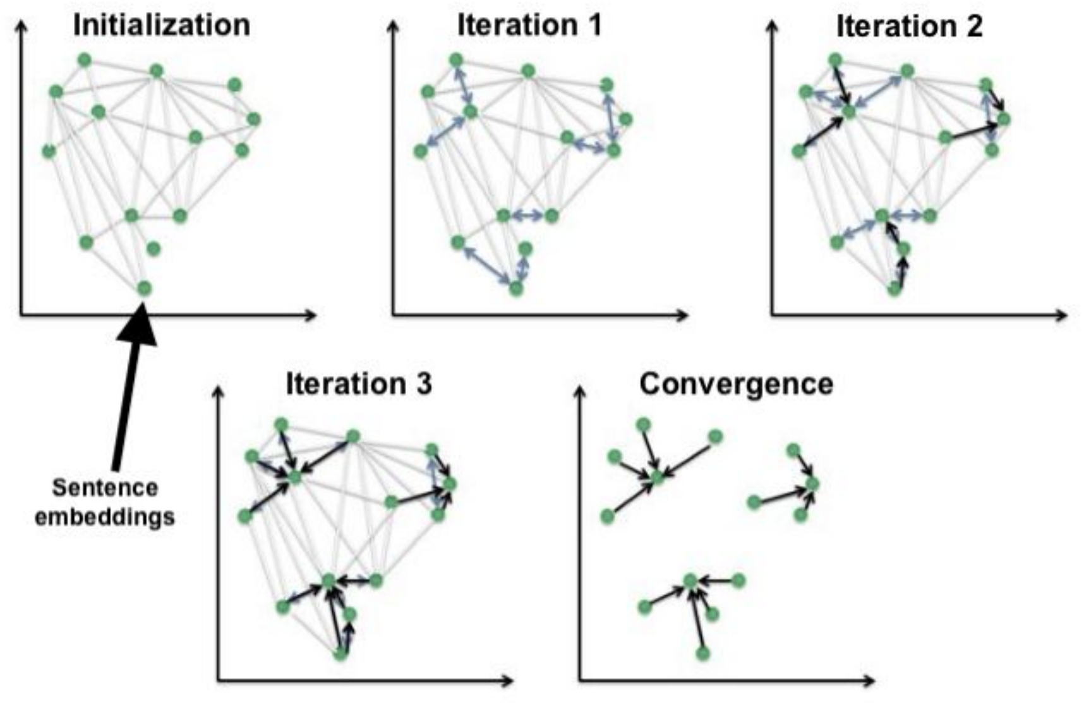
- Measuring unsupervised learning is challenging
- We propose the following evaluation techniques:
  - Automated: ROUGE-1, semantic similarity, sentiment accuracy
  - By hand: human evaluation, consistency preservation

ROUGE-1 (out of 100)	Semantic Similarity	Sentiment Accuracy	Human Eval (out of 10)	Content Preservation
Word overlap between summary and the aggregate reviews	Cosine similarity of the mean of the sentence embeddings between summary and aggregate reviews	Trained sentiment classifier on 100,000 reviews with simplified 3 categories. 80% accuracy on the test set. Compare sentiment of summary versus average sentiment of reviews.	For each of the 5 sentences in the generated summary. Award 2 points if it matches the summary done by a human or if the point is valid.	For each of the 5 sentences in the generated summary, we grade the question "Does the content of summary represent the most commonly described features in cluster review sentences?" on a 1-5 Likert scale.

- Baseline: K-means with 5 clusters

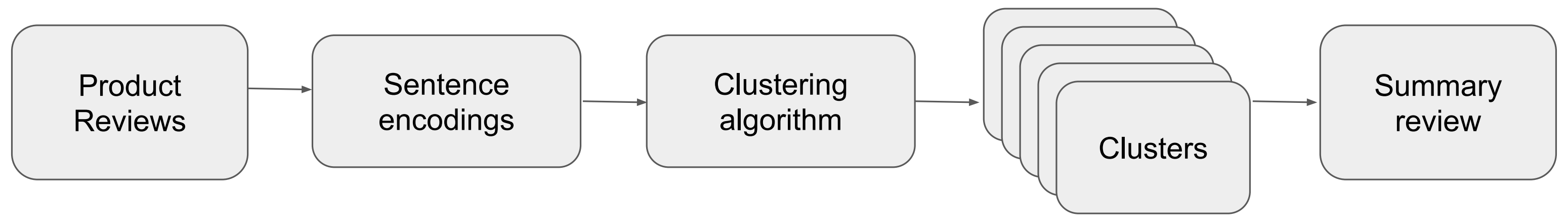
Extractive Summarization

- We cluster or rank sentences using their dense vector representations
- Sentences are picked from each cluster to create summary.
- **Algorithms:**
  - *k-means*
  - *Affinity propagation*
  - *DBSCAN*
  - *PageRank*



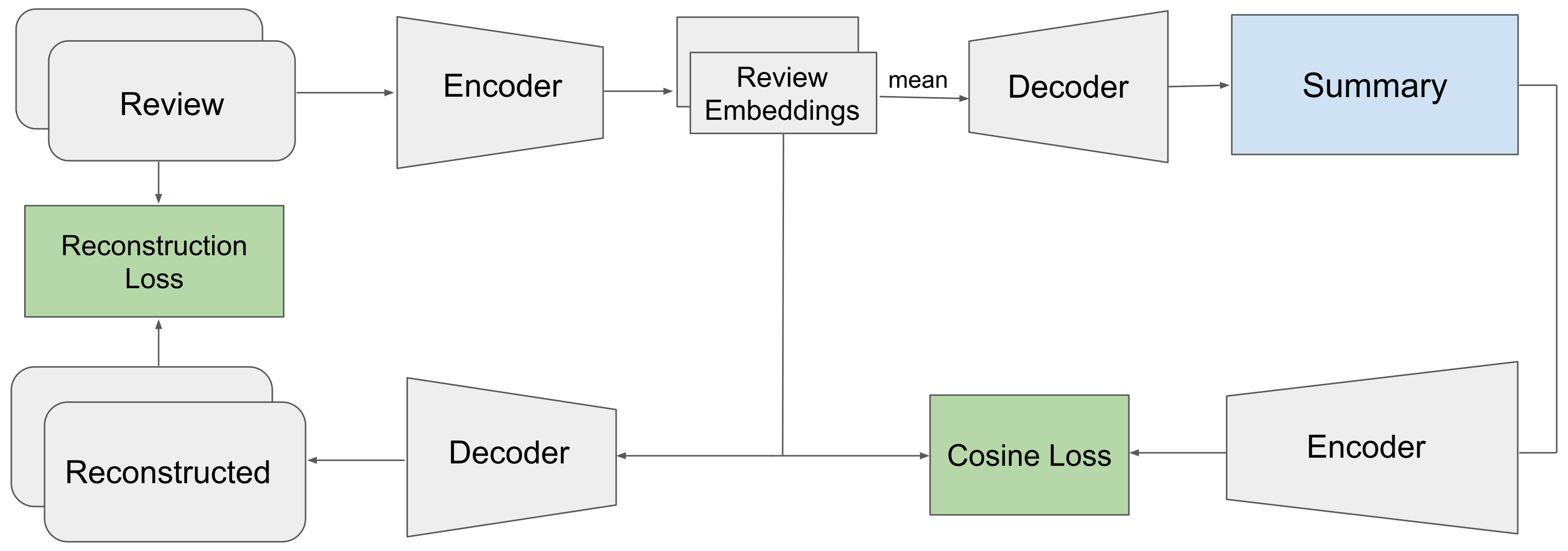
Affinity propagation

- **Details:**
  - K-means requires pre-set cluster count, others find automatically
  - Sentence encoders used: NNLM, USE, Word2Vec [2]



Abstractive Summarization

- There are two parts to this model [3]
  - A sequence to sequence autoencoder
  - A natural language summary generator



- The sequence to sequence autoencoder is trained on trying to encode every review and reconstruct it using cross entropy loss
- The cosine loss tries to minimize the cosine distance between the encoded summary and the encoding of every single review
- Encoder and decoder weights are tied on both halves of the model

Final Results and Evaluation

Model	ROUGE-1 (out of 100)	Semantic Similarity (out of 100)	Sentiment Accuracy	Human Eval (out of 10)	Content Preservation
PageRank	21.7	96.3	0.64	6.67	3.93
Affinity	22.1	94.3	0.64	4	3.67
DBScan	16.7	79.9	0.63	2	3.85
Random	18.5	97.6	0.57	2	2.33
Baseline: Kmeans	21.3	93.6	0.66	4	3.44

Qualitative Results (Affinity)

#	Theme Sentence	Mentioned By
1	I wasn't able to take a picture until the next day.	24 People
2	Have owned this camera for a few years.	22 People
3	I got this camera a couple of months ago and I'm not real please with it.	24 People
4	The camera has a good zoom on it, and is very easy to use.	17 People
5	If you need small cameras, you have to typically settle for picture quality that LOOKS like it came from a tourist gadget.Not this one!	15 People

Example of summarization given by Affinity Propagation. This summary scored a 4/10 on human eval as 2 of the mentioned themes are insightful (Green), and 3 of the themes do not provide much information (Red). PageRank algorithm gave better summaries, though do not give counts on how frequently the theme is mentioned.

Challenges and Error Analysis

- ***Clustering errors:*** unrepresentative themes or theme crossover (m)
- ***Attribute match:*** unrepresentative sentiment applied to feature (r)
- ***Content preservation:*** main theme of cluster not captured (r)
- ***Conciseness:*** extractive summary sentences can often refer to multiple things (c)
  - E.g. *"So for the burner it's 5 stars - maybe the software will work with…"*
- ***Out-of-context errors:*** sentences pulled extractively may appear out of context when pulled from a multi-sentence review (r)
  - E.g. *"Unlike Kodak, which has provided me with 4 coasters out of 15 used."*
- ***Abstractive summarization:*** *pending results due to long training*

Error frequency: common (c), moderate (m), rare (r)

References

1. J. J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-based recommendations on styles and substitutes. *SIGIR,* 2015
2. Token Based Text Embedding Trained on English Google News 200B Corpus. TensorFlow Hub, 2018, tfhub.dev/google/nnlm-en-dim128-with-normalization/1., https://tfhub.dev/google/universal-sentence-encoder-large/3.
3. Chu, E., & Liu, P. J. Unsupervised Neural Multi-document Abstractive Summarization. arXiv preprint arXiv:1810.05739, 2018