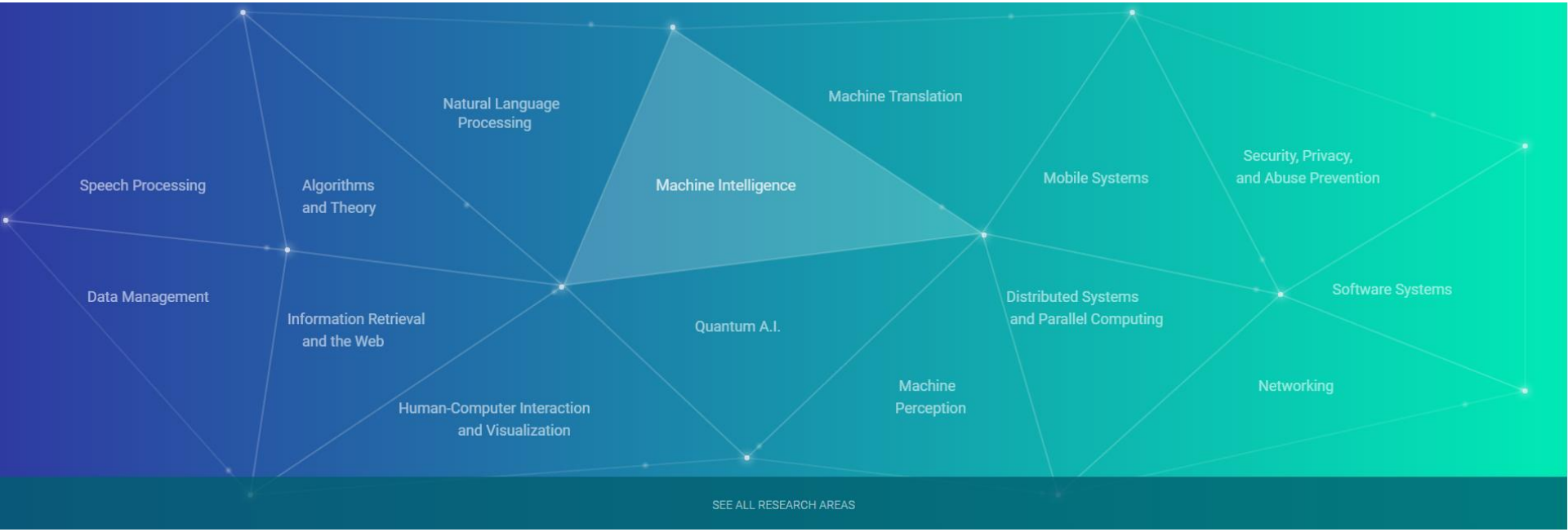
# Paper Review

Smart reply : automated response suggestion for email

davinnovation@gmail.com

# Research at Google

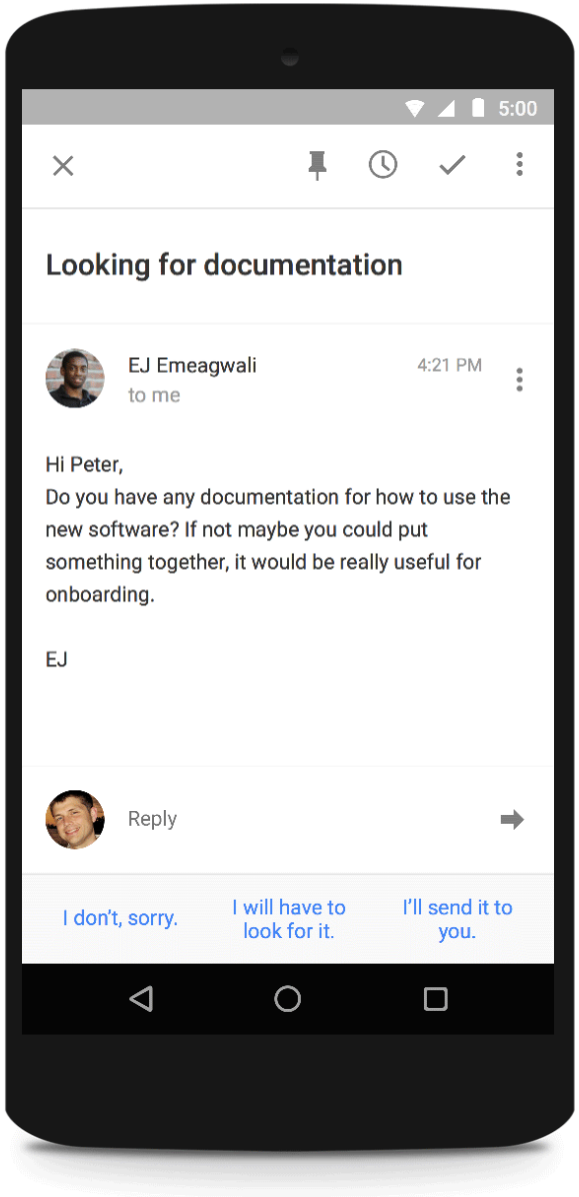



Google research program … for faculty and others

# Smart Reply : automated response suggestion for email 2016. 06. 15

2015년 하반기에 추가된 메일에 '짧은' 답장 추천
Inbox, Gmai 의 기능

모바일 환경에서 10% 비율의 답장을 도와줌




Keyword : LSTM; Deep Learning; Clustering; Semantics;

# Motivation

**Email**은 웹에서 가장 널리 사용되는 커뮤니케이션 도구!

Social Network 사용자가 늘어나고 있으나, **수십억의 인구**가 지속적인 사용 중

사용자가 메시지에 답장을 하는 것이 'challenging'하다. 모바일에서 입력하는 것은 시간 낭비가 될 수도 있고, **25%의 답장이 20개 이하 단어**로 이루어져 있음

# Goal...

**Response Quality** 항상 높은 품질의 개인화된 답장

**Utility** 여러 답장 중 하나라도 사용자가 선택할 수 있게

**Scalability** 수 많은 메시지를 지연 없이 효과적으로 처리

**Privacy** 요약 통계를 제외한 데이터 검사 없는 시스템

# Process

Input : incoming message

Output : possible replies

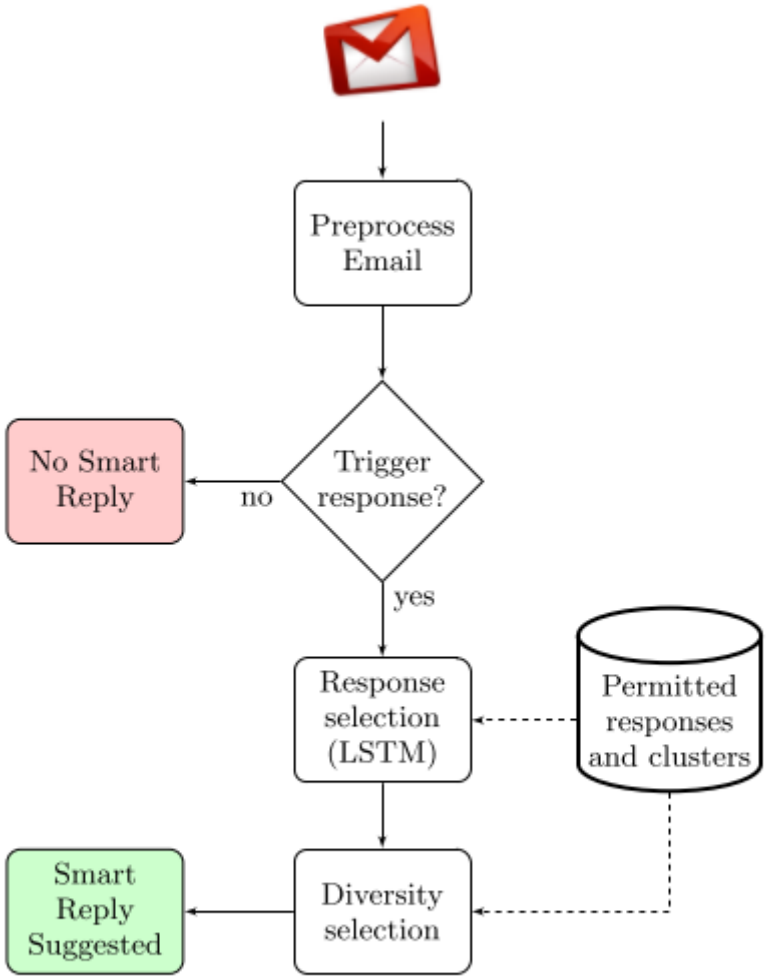
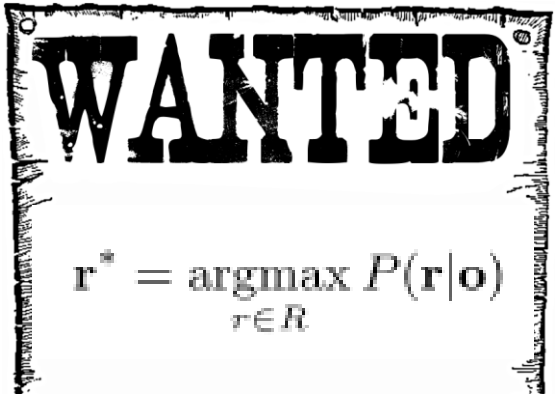



**Figure 2: Life of a message. The figure presents the overview of inference.**

# Selecting Responses

R : all possible response

o : given original message




LSTM Model : sequence to sequence learning [http://arxiv.org/abs/1409.3215]

r : sequence of tokens (conditional probability of response tokens)

o : sequence of tokens(original message)

$$P(r_1, ..., r_m | o_1, ..., o_n) = \prod_{i=1}^{m} P(r_i | o_1, ..., o_n, r_1, ..., r_{i-1})$$

is interpreted as $P(r_t | o_1, ..., o_n, r_1, ..., r_{t-1})$. Given the factorization above, these softmaxes can be used to compute $P(r_1, ..., r_m | o_1, ..., o_n)$.

- Training : AdaGrad [stochastic gradient descent]

$$G = \sum_{\tau=1}^{t} g_\tau g_\tau^{\mathsf{T}}$$

where $g_\tau = \nabla Q_i(w)$, the gradient, at iteration $\tau$. The diagonal is given by

$$G_{j,j} = \sum_{\tau=1}^{t} g_{\tau,j}^2.$$

This vector is updated after every iteration. The formula for an update is now

$$w := w - \eta \operatorname{diag}(G)^{-\frac{1}{2}} \circ g^{[a]}$$

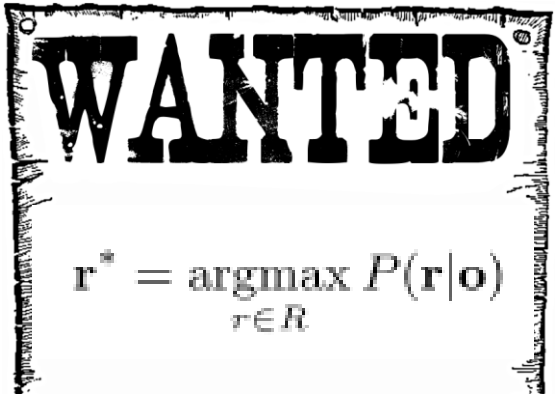
or, written as per-parameter updates,

$$w_j := w_j - \frac{\eta}{\sqrt{G_{j,j}}} g_j.$$

# Selecting Responses ‘

R : all possible response

o : given original message




| Query | Top generated responses |
|---|---|
| Hi, I thought it would be great for us to sit down and chat. I am free Tuesday and Wenesday. Can you do either of those days?<br><br>Thanks!<br><br>–Alice | I can do Tuesday.<br>I can do Wednesday.<br>How about Tuesday?<br>I can do Tuesday!<br>I can do Tuesday. What time works for you?<br>I can do Wednesday!<br>I can do Tuesday or Wednesday.<br>How about Wednesday?<br>I can do Wednesday. What time works for you?<br>I can do either. |

# Challenges

## Response Quality Problem

- Poor grammar, spelling…. (your the best!)
- Too informal (yup, got it thx)
- Offensive (Leave me alone)

=> Construct response space **R**

| Unnormalized Responses | Normalized Responses |
|---|---|
| Yes, I'll be there. | Sure, I'll be there. |
| Yes, I will be there. | Yes, I can. |
| I'll be there. | Yes, I can be there. |
| Yes, I can. | Yes, I'll be there. |
| What time? | Sure, I can be there. |
| I'll be there! | Yeah, I can. |
| I will be there. | Yeah, I'll be there. |
| Sure, I'll be there. | Sure, I can. |
| Yes, I can be there. | Yes. I can. |
| Yes! | Yes, I will be there. |
| **Normalized Negative Responses** | |
| Sorry, I won't be able to make it tomorrow. | |
| Unfortunately I can't. | |
| Sorry, I won't be able to join you. | |
| Sorry, I can't make it tomorrow. | |
| No, I can't. | |
| Sorry, I won't be able to make it today. | |
| Sorry, I can't. | |
| I will not be available tomorrow. | |
| I won't be available tomorrow. | |
| Unfortunately, I can't. | |
| **Final Suggestions** | |
| Sure, I'll be there. | |
| Yes, I can. | |
| Sorry, I won't be able to make it tomorrow. | |

**Table 2: Different response rankings for the message *"Can you join tomorrow's meeting?"***

## Utility Problem

- Little diversity

=> Light normalization, Suggestion Diversity

| Query | Top generated responses |
|---|---|
| Hi, I thought it would be great for us to sit down and chat. I am free Tuesday and Wenesday. Can you do either of those days?<br><br>Thanks!<br><br>–Alice | I can do Tuesday.<br>I can do Wednesday.<br>How about Tuesday?<br>I can do Tuesday!<br>I can do Tuesday. What time works for you?<br>I can do Wednesday!<br>I can do Tuesday or Wednesday.<br>How about Wednesday? |

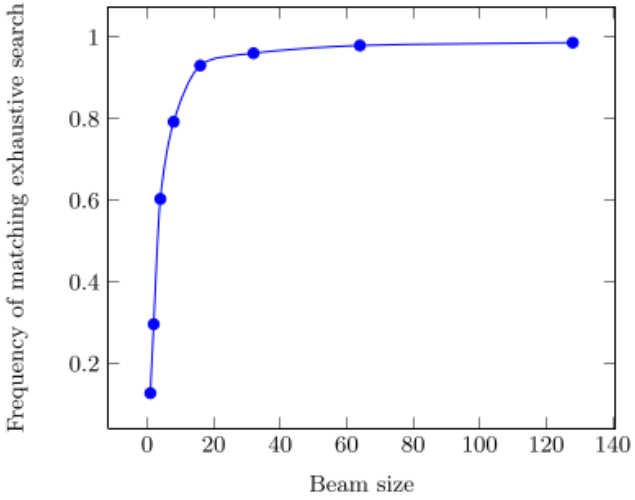
# Challenges

## Scalability Problem

- R set is very large
- need ASAP!

## => Left to Right beam search

휴리스틱 탐색 기법의 하나로 탐색 도중에 해가 되는 후보 가지가 여러 개 있을 때 해가 될 가능성이 큰 가지만을 남기고 나머지는 모두 잘라 버리는 방식




Beam 사이즈가 16만 되어도 brutal force와 93% 유사한 정답을 냄

## Privacy Problem

- Encrypted! -> only frequent words can be accessed && statistics on anonymized sentence

# Response Set Generation

## Set generation == satisfy(response quality, utility)

(Yes, I'll be there == I will be there).consider as same

1. Canonicalizing(normalizing) email response
   - Modifiers || unattached to head words are ignored from sentence
2. Semantic intent clustering
   - 'Thank you' vs 'sorry' vs 'cannot make it'
   - "Ha ha", "lol", "Oh that's funny" : funny cluster
3. Graph construction
   - Add seed (thanks -> "Thanks!", "Thank you") : test 클러스터 100개에 3-5개의 seed 단어 설정
   - Frequent response message as node (Vr) : (Thanks, I love you, sounds good)
   - Lexical(grammatical) features as node (Vf)
   - (Vr, Vf) edge ➔ make manua;ly labeled example (Vl)

   Observation : (Let us get together soon, When should we met?) && (When should we met?, How about Friday?) : response used to question
4. Semi-supervised learning

# Response Set Generation ’

## Set generation == satisfy(response quality, utility)

(Yes, I’ll be there == I will be there).consider as same

4. Semi-supervised learning

- semantic labeling for all response=> used EXPANDER

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 807–814, 2010

minimize

$$s_i||\hat{C}_i - C_i||^2 + \mu_{pp}||\hat{C}_i - U||^2 + \mu_{np}\Big(\sum_{j\in\mathcal{N}_{\mathcal{F}}(i)} w_{ij}||\hat{C}_i - \hat{C}_j||^2 + \sum_{j\in\mathcal{N}_{\mathcal{R}}(i)} w_{ik}||\hat{C}_i - \hat{C}_k||^2\Big) \quad (1)$$

s == [0,1] // node i가 seed면 1

C // node i의 learned semantic cluster distribution

Nf, Nr // node i의 이웃

뮤np // predefined penalty

뮤pp // penalty for label distribution deviating from the prior

U // uniform distribution

If no seed $\mu_{np}\sum_{i\in\mathcal{N}(j)} w_{ij}||\hat{C}_j - \hat{C}_i||^2 + \mu_{pp}||\hat{C}_j - U||^2 \quad (2)$

# Response Set Generation ’’

Set generation == satisfy(response quality, utility)

(Yes, I’ll be there == I will be there).consider as same

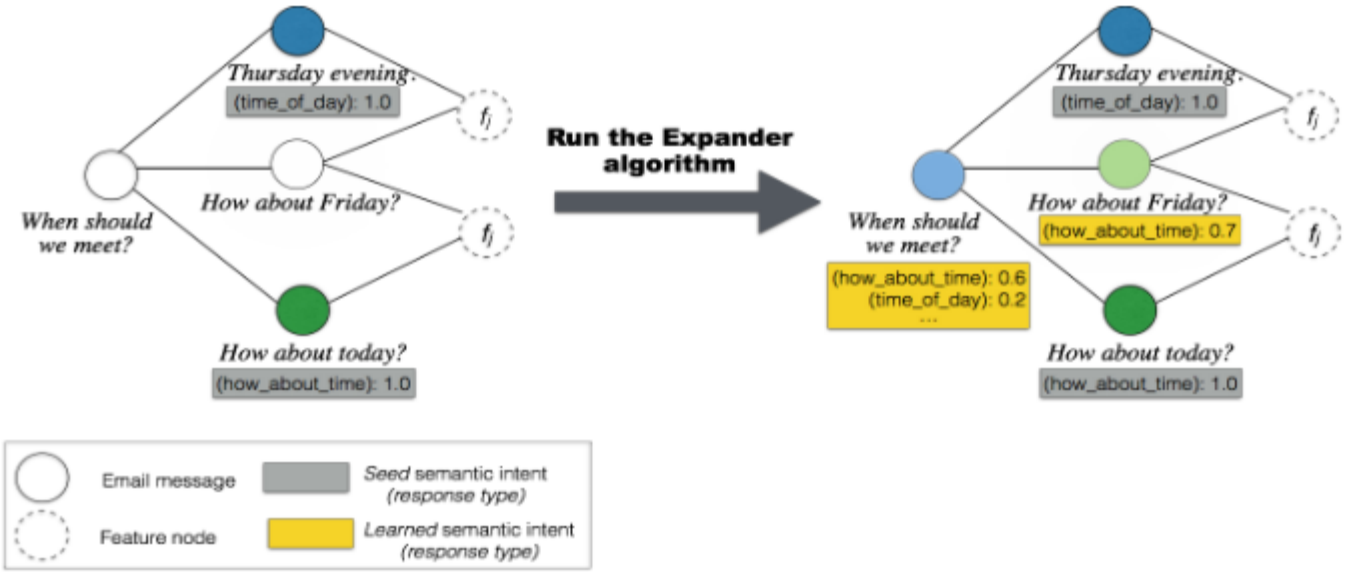



**Figure 4: Semantic clustering of response messages.**

Validation : Top response chosen!

# Suggestion diversity

We need to choose a small number of options to choose!

| Unnormalized Responses | Normalized Responses |
|---|---|
| Yes, I'll be there. | Sure, I'll be there. |
| Yes, I will be there. | Yes, I can. |
| I'll be there. | Yes, I can be there. |
| Yes, I can. | Yes, I'll be there. |
| What time? | Sure, I can be there. |
| I'll be there! | Yeah, I can. |
| I will be there. | Yeah, I'll be there. |
| Sure, I'll be there. | Sure, I can. |
| Yes, I can be there. | Yes. I can. |
| Yes! | Yes, I will be there. |
| **Normalized Negative Responses** | |
| Sorry, I won't be able to make it tomorrow. | |
| Unfortunately I can't. | |
| Sorry, I won't be able to join you. | |
| Sorry, I can't make it tomorrow. | |
| No, I can't. | |
| Sorry, I won't be able to make it today. | |
| Sorry, I can't. | |
| I will not be available tomorrow. | |
| I won't be available tomorrow. | |
| Unfortunately, I can't. | |
| **Final Suggestions** | |
| Sure, I'll be there. | |
| Yes, I can. | |
| Sorry, I won't be able to make it tomorrow. | |

**Table 2: Different response rankings for the message *"Can you join tomorrow's meeting?"***

Strategy

1. Omitting redundant response
   -> Make Response set R
2. Enforcing negative || positive response
   -> 점수를 매겨보니 부정적인 답은 전체적으로 적은 점수를 얻음
   -> 2개는 긍정 하나는 부정으로 하자!

# Triggering

Entry point of Smart Replay System

Currently, system decides to response 11%

1. “Where do you want to go today?” -> 추천 필요 없음
   - 필요한 것만 추천하자
2. Fast triggering

# Triggering ‘

## Entry point of Smart Replay System

## Training set

(incoming message, [true, false]) : true는 모바일에서 답장 된 것
-> true가 된 incoming message를 body, subject, headers로 분리 + address book 존재여부 + 답장한 적이 있는지 여부

## Architecture

Feedforward multilayer perceptron with embedding layer && three fully connected hidden layers
Activation function : ReLu
Dropout : applied after each hidden layer
Trained with : AdaGrad

# Evaluation && Results

## Used Data

- Language detection
- Tokenization
- Sentence segmentation
- Normalization
- Quotation removal
- Salutation/close removal (like *Best regards, Mary*)

=> 2.38억 messages ( 1.53억 messages는 no response)

# Evaluation && Results

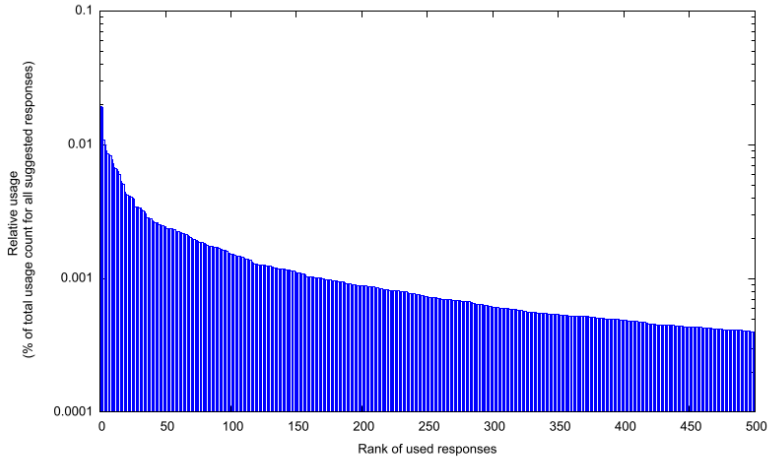
Triggering : 11%가 답장이 가도록 됨

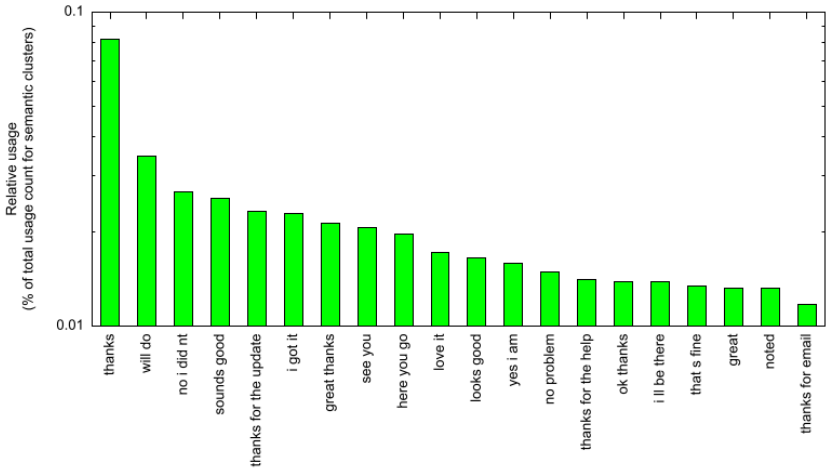
Message response : 많은 새로운 cluster가 생성되는 중.
하지만 답장 제안이 만들어져도 스크롤을 끝까지 안내리거나, web을 사용하는 것을 알 수 있다.
Smart Reply 사용자 중
45%는 첫번째 답장, 35%는 2번째, 나머지는 3번째의 답장을 선택함

단순히 가장 높은 3개의 답만 보여주니 선택률이 7.5% 줄었다.

| | Daily Count | Seen | Used |
|---|---|---|---|
| Unique Clusters | 376 | 97.1% | 83.2% |
| Unique Suggestions | 12.9k | 78% | 31.9% |

**Table 4: Unique cluster/suggestions usage per day**



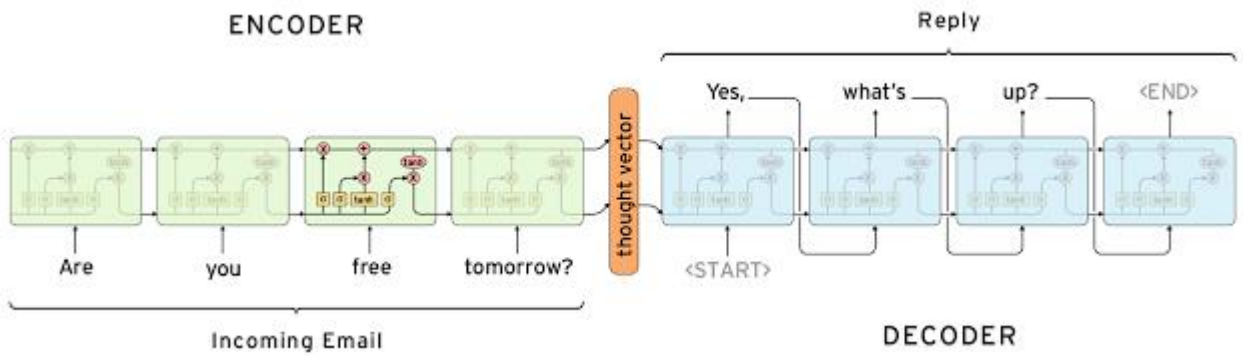
# Reference

https://research.googleblog.com/2015/11/computer-respond-to-this-email.html

Smart Reply: Automated Response Suggestion for Email

Didn't read but important

http://arxiv.org/abs/1409.3215 [ sequence to sequence learning ]




Thought vector : Squeeze to similar
for example, the vector for "Are you free tomorrow?" should be similar to the vector for "Does tomorrow work for you?"