Disco Sumo

Sander Wubben, Suzan Verberne, Antal van den Bosch
and Emiel Krahmer

Discussion Thread Summarisation for Mobile Devices

- Many people look for answers on forums
- Often using mobile devices
- Problems:
  - Information not easily accessible
  - Conflicting opinions
  - Intention of searcher not clear

How many cats is too many

Sander

https://www.reddit.com/r/AskReddit/comments/3oq025/how_many_cats_is_too_many/

askreddit

Want to join? Log in or sign up in seconds.

COMMENTS RELATED

11

How many cats is too many? (self.AskReddit)

submitted 2 hours ago by PhillipCarey

55 comments share

Ask A New Question

[SERIOUS]

search

all 55 comments

sorted by: best

[-] smokeycat76 14 points 2 hours ago

If you can't properly clean their litter box(es), feed, and play with them, you have too many. For some people one is too many.

permalink

[-] ancisfranderson 5 points 2 hours ago

If you're asking, it's already too late.

permalink

[-] MitchNYM 4 points 2 hours ago

I have three. I say more than that is too many.

permalink

[-] _CHILLBRO_SWAGGINS 1 point an hour ago

I agree. 3 is pushing it and 4 is just too many cats

permalink parent

[-] gnrl2 2 points an hour ago

5 is right out.

permalink parent

Submitted on 14 Oct 2015

11

(72% upvoted)

https://redd.it/3oq025

username password

remember me reset password login

Welcome to /r/AskReddit

unsubscribe

9,800,494 subscribers 42,377 online now

New mod tools by: - December 31

The admins have agreed to better communication with mods and to release improved mod tools by

How many cats is too many

https://www.reddit.com/r/AskReddit/comments/3oq025/how_many_cats_is_too_many/

[-] **thursdaycookies** 2 points 2 hours ago

I tend to use an equation: amount of bedrooms - 1 = max amount of cats allowed in that home. This assumes, however that you have enough space and money for their food, litter, and toys.

permalink

[-] **UrukHaiGuyz** 2 points 2 hours ago

I'd say one 1.5 cats per adult in a household max.

permalink **parent**

[-] **Hullabalooga** 2 points 2 hours ago

If you're asking this question, you already have too many.

permalink

[-] **itsamee** [score hidden] 54 minutes ago

1-2: Perfectly normal

3-4: Wouldn't do it myself but still acceptable

5-7: Damn, you must really love cats!

8-12: Holy shit!!

13+: Calling animal protection on you!

permalink

[-] **Hybe529** 1 point 2 hours ago

Depends. If I had a farm I'd "have" like 15 of those little beasts.

permalink

[-] **Vagfilla** 1 point 2 hours ago

We have four. Four is pushing it but manageable. They do go outside so the boxes aren't as bad as they could be.

permalink

6 Questions seeking professional advice are inappropriate for this subreddit and will be removed. more >>

7 Soliciting money, goods, services, or favours is not allowed. more >>

8 Mods reserve the right to remove content or restrict users' posting privileges as necessary if it is deemed detrimental to the subreddit or to the experience of others. more >>

9 Comment replies consisting solely of images will be removed. more >>

If you think your post has disappeared, see spam or an inappropriate post, please do not hesitate to contact the mods, we're happy to help.

Tags to use:

[Serious]

Use a **[Serious]** post tag to designate your post as a serious, on-topic-only thread.

Filter posts by subject:

Mod posts | Serious posts | Megathread | Breaking news

Problem statement

How can opinionated information in discussion threads be summarized for mobile devices?

- how to distinguish relevant and important information from irrelevant or redundant information?
- how can we use the specific structure of time-stamped interactions between messages in a thread to distinguish relevant and important information?
- how can sentiments in contributions to discussion threads be determined and aggregated?
- how can all obtained information be integrated into an overall summary?

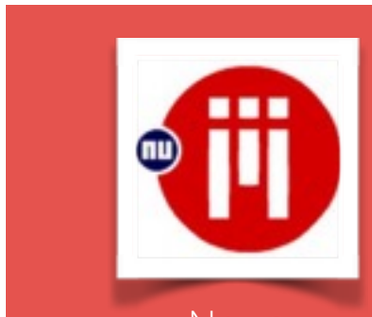
Data

News

?

English

★



News

90K

Dutch

★

Personal interest

3.5M

English

★

Personal interest

1.5M

Dutch

Data

1.65 billion

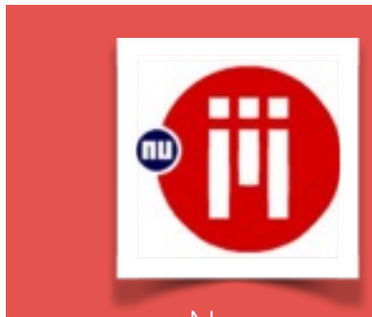
comments

News

?

English

★



News

90K

Dutch

★

Personal interest

3.5M

English

★

Personal interest

1.5M

Dutch

Dataset

- 1.65B reddit comments
- Collected from october 2012 - august 2015
- Comment are grouped by month
- Challenges:
  - reconstruct comment threads
  - link to opening post
- Which threads do we use?
  - /r/nostupidquestions
  - /r/askreddit
  - /r/askmen
  - /r/askwomen
  - /r/myfriendwantstoknow
  - /r/answers
  - ....

Recognizing the type and structure of the thread

- The type of thread should determine the summarization strategy.
- If a closed question is asked (e.g. “ms office to libre/open office?” on reddit)

  count votes

  present the counts with the most important arguments
- If the topic is more complex (e.g. the Guardian article “Turkey says Kurdish peace process impossible as Nato meets”)

  summarize different points of view
- If the topic is very personal (e.g. “I’m 28 years old, I have breast cancer. Here’s my story.” on reddit)

  summarizing the thread may not be helpful at all for the interested reader.

Document type definition (DTD) for forum threads.

<!ELEMENT thread (threadid,post+,category*,type*,nrofviews?)>

<!ELEMENT post (postid,author,timestamp, parent*,upvotes?,downvotes?,body)>

<!ELEMENT author (#PCDATA)>

<!ELEMENT timestamp (#PCDATA)>

<!ELEMENT parent (#PCDATA)>

<!ELEMENT upvote (#PCDATA)>

<!ELEMENT downvote (#PCDATA)>

<!ELEMENT body (content,url*)>

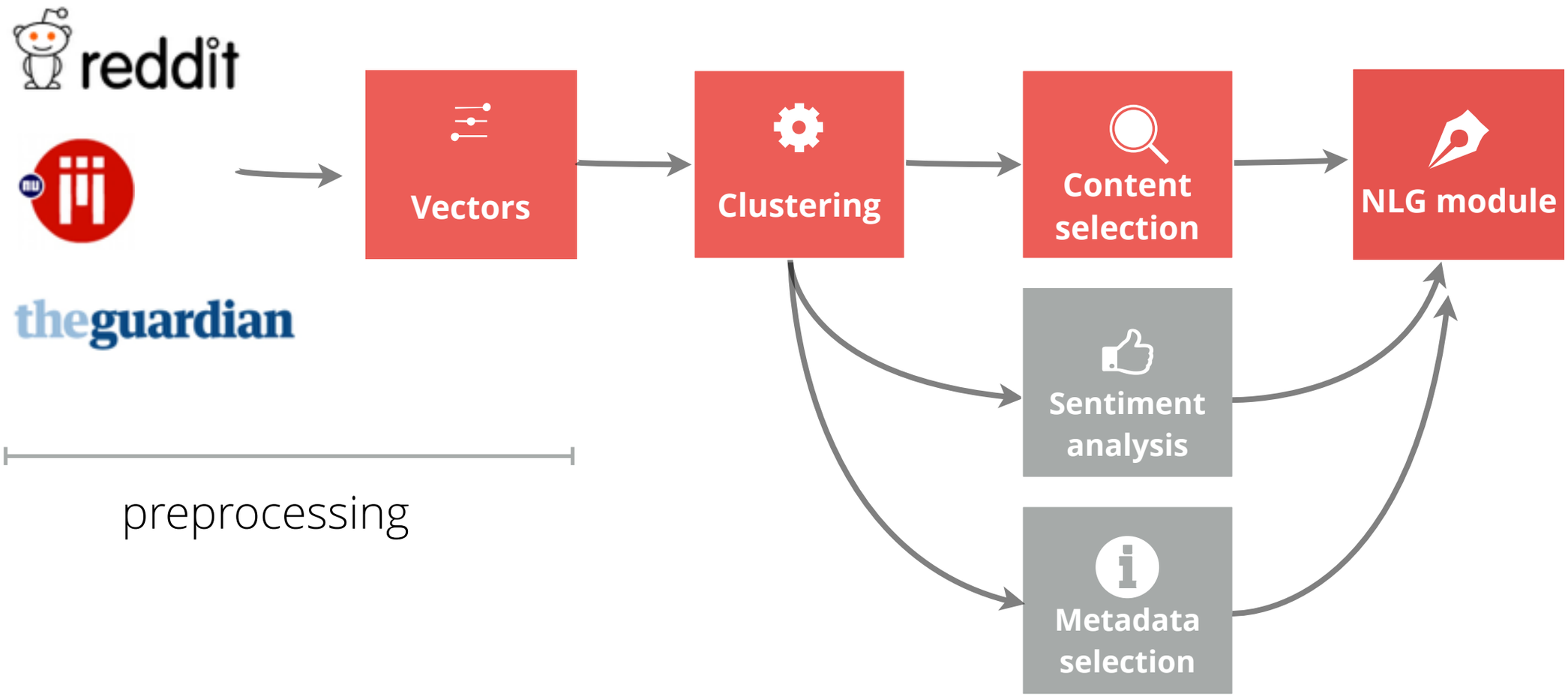
<!ELEMENT content (#PCDATA)>

<!ELEMENT url (#PCDATA)>

General system design

- Make the system highly modular
- Aim for domain independence
- Aim for language independence for main modules
- Where possible: unsupervised
- Minimize need for feature engineering/ annotated data (be able to make use of crowd annotated data)
  upvotes, best answer etc.

System design



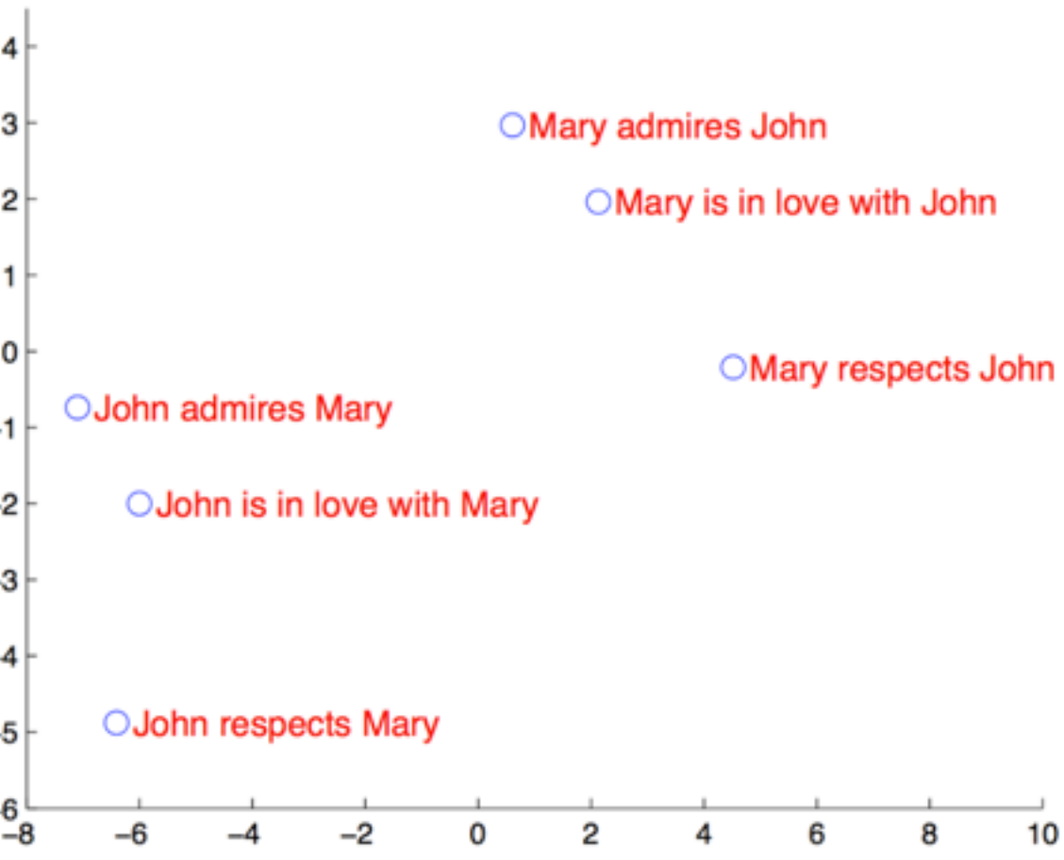
preprocessing

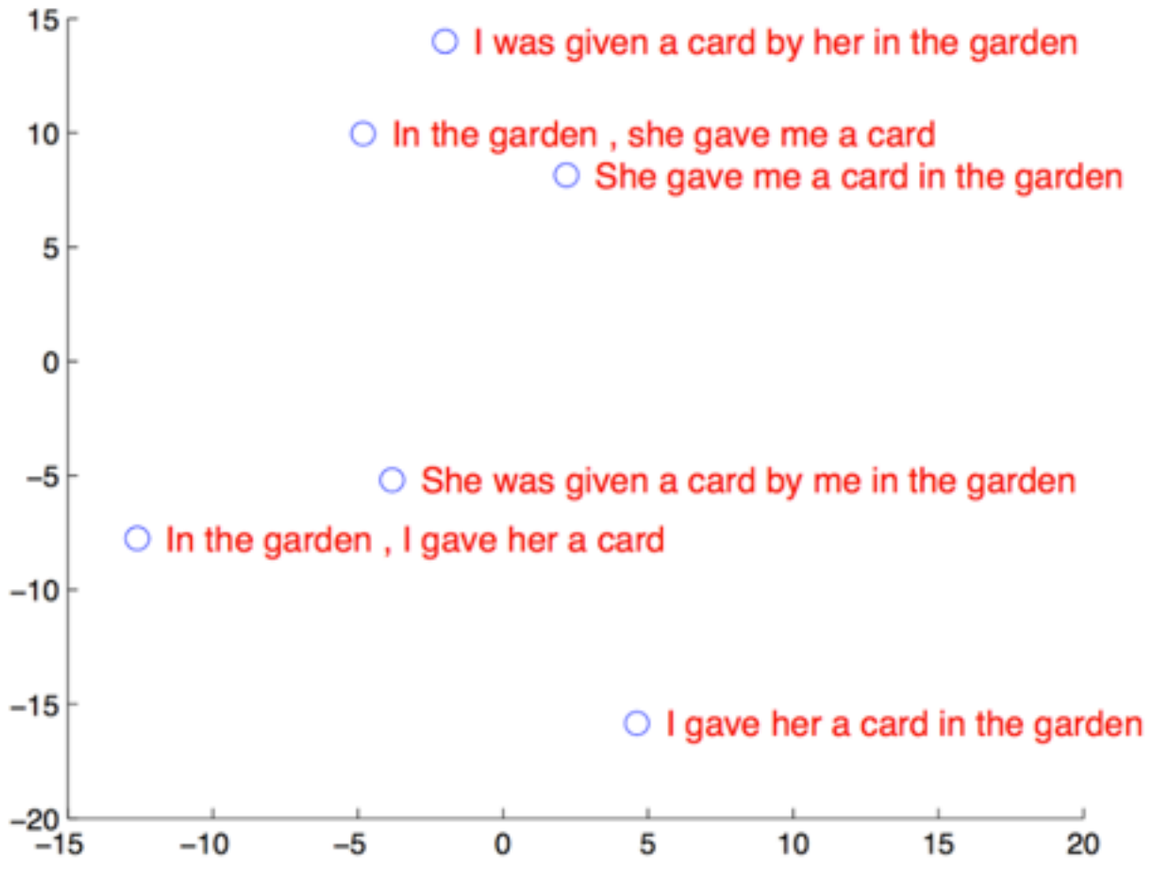
Preprocessing

- **Robust, domain/language independent system:**
  create vector representations for sentences/posts
- **Learn word embeddings**
  Distributional vector representation of words (Mikolov et al, 2013)
- **Combine word embeddings into sentence/paragraph/post vectors**
  Paragraph vectors (Quoc and Mikolov, 2014)
  Skip-thought vectors (Kiros et al, 2015)

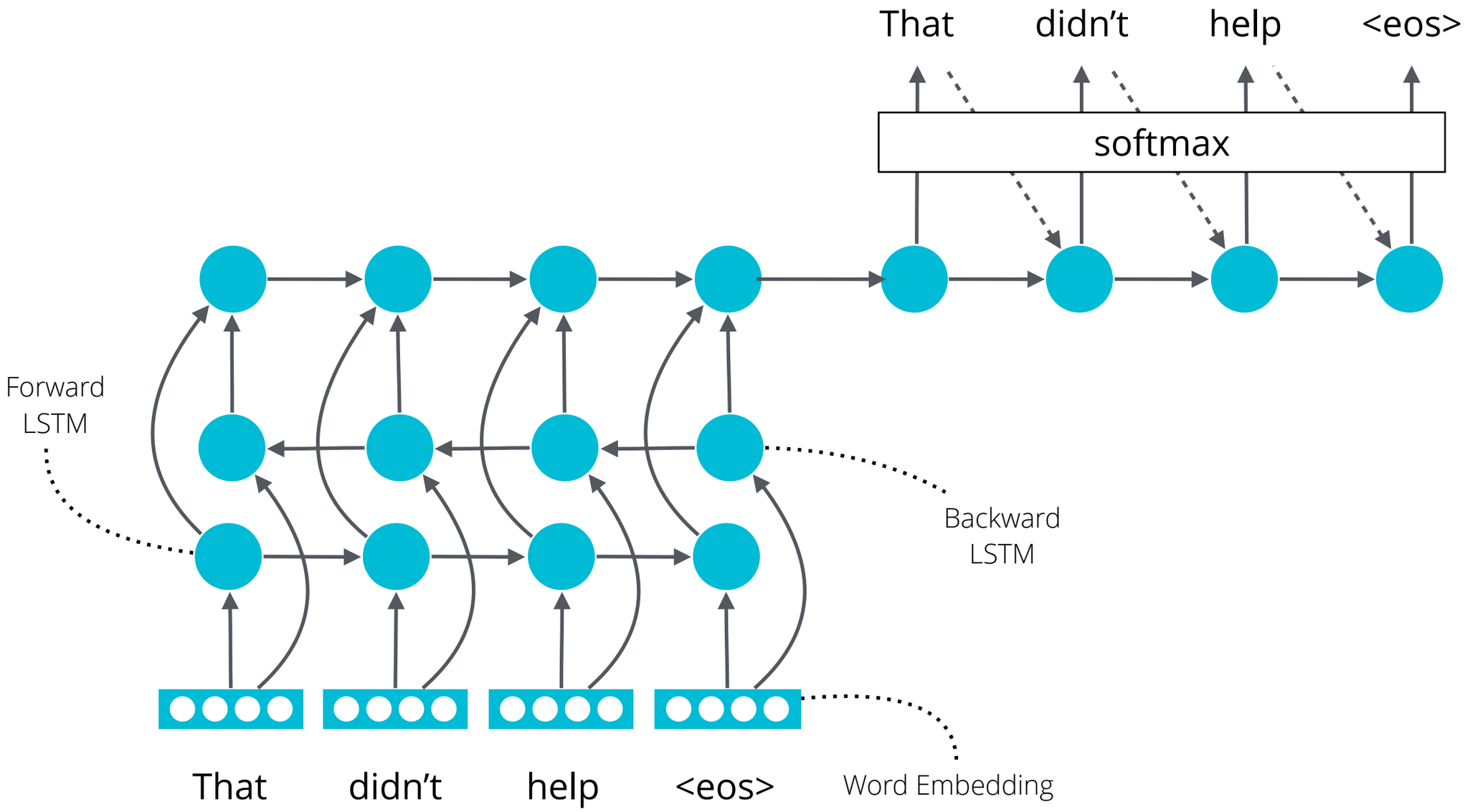
Clustering with sentence vectors

(Sutskever et al, 2014)

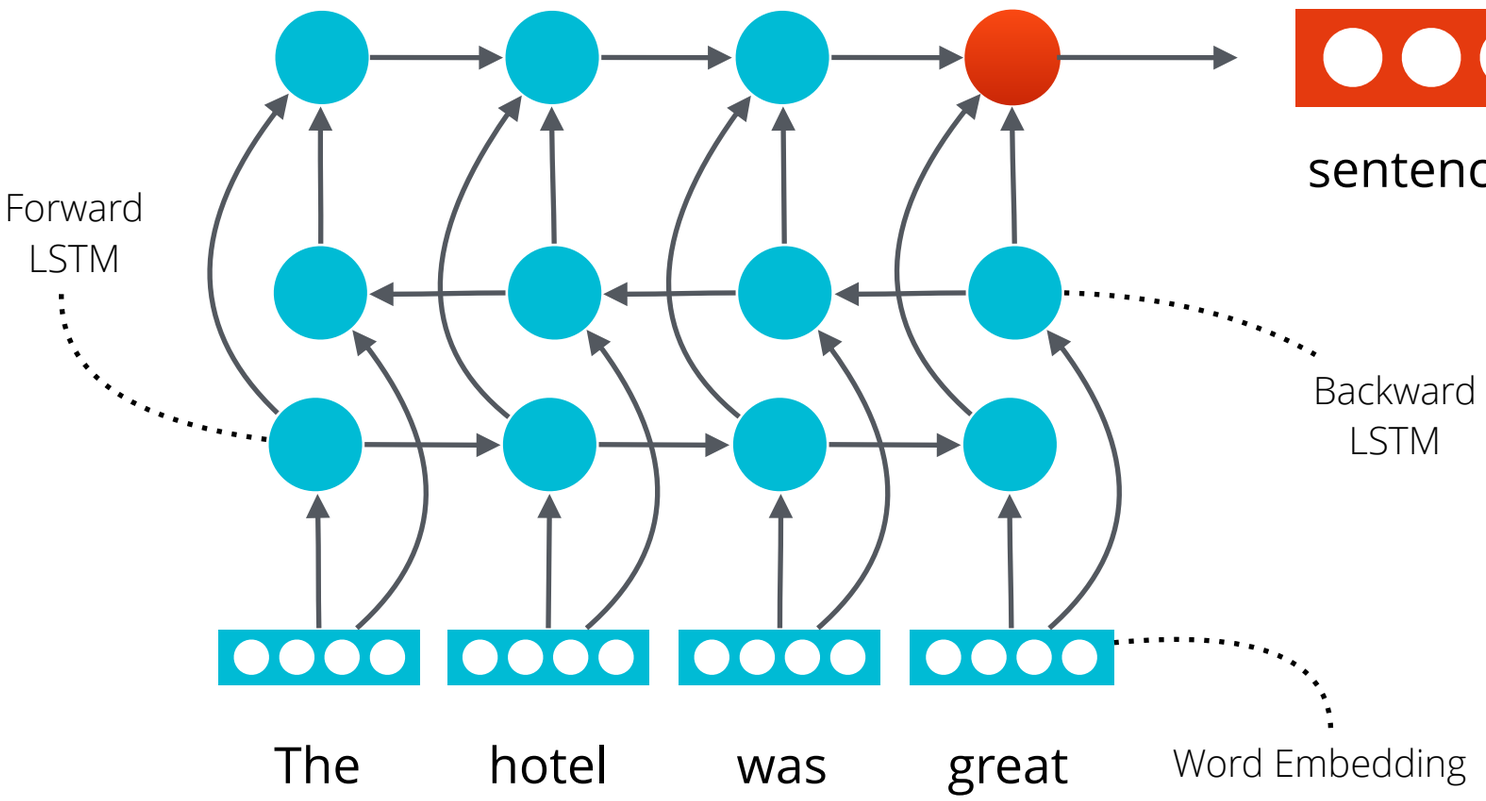




Encoder/Decoder



Encoder


sentence vector
Forward LSTM
Backward LSTM
The
hotel
was
great
Word Embedding

The spirit of reddit…

During training, the autoencoder makes ‘safe’ predictions:

the nature of man is brutish , greedy , irrational , and mean .

- original

the theory of reddit is born , islamic , unhealthy , and mean

- generated

First intuitions

- Reconstructing source sentence works well
- …but extracted vectors don’t seem to capture semantics
- Possibly train on harder seq2seq task

Reply prediction (similar to (Vinyals, 2015))

Machine Translation

Paraphrasing

TO BE CONTINUED...

PhD student vacancy

- Statistical Natural Language Generation for Robot Journalism
- Soon available!

Questions?