What to ask Questions about?

Sören Etler
Universität Potsdam
etler@uni-potsdam.de

Motivation

Identifying question-worthy tokens

- Finding bias in the data set
- First step in a question generation pipeline
- Finding important information / concepts

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Evaluation of QA models

- Identifying weaknesses of current QA models
- Compare the performance of three high ranking models
  - BERT (Google AI Language)
  - BiDAF + Self Attention + ELMo (Allen Institute for AI)
  - nlnet (Microsoft Research Asia)

**Can tokens that are very likely to be the answer to a potential question be predicted (question-worthy tokens)?
Does this have an impact on the performance of QA systems?**

The Stanford Question Answering Dataset (SquAD 2.0)

- 150,000 questions about 19,000 paragraphs
- Created my Crowdworkers on Amazon Mechanical Turk
- Answers sequences of answerable & unanswerable questions are marked as question-worthy tokens

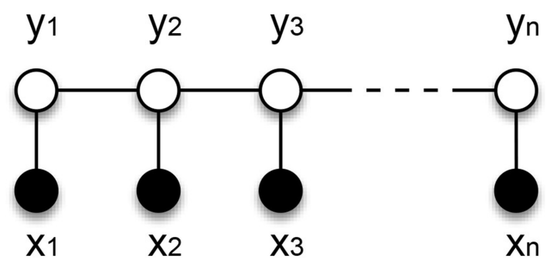
Overview of the data set

	Training data	Test data
Texts / paragraphs	442 / 19,035	35 / 1,204
Question-worthy tokens (I)	351,862	35,214
Non-question-worthy tokens (0)	2,244,574	141,594
Mean paragraph length	136	147

Identifying question-worthy tokens

Model

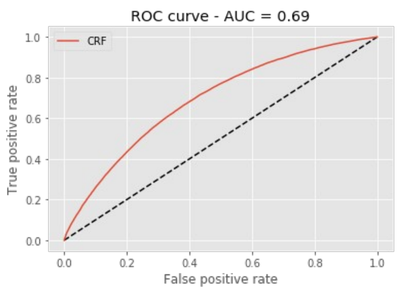
- Conditional Random Field (CRF)
  - implemented using sklearn_crfsuite
  - cross validation of hyperparameters on random subset
- Features
  - Lemma of the words
  - POS Tags
  - Named Entities
  - Dependencies (ClearNLP Dependency Labels)
  - Stopwords
  - Position in the Text



Evaluation

- Evaluation metrics
  - Precision, Recall, F1-Score
  - ROC-Curve (AUC)
  - Log Loss

	precision	recall	f1-score	support
I	0.53465	0.00153	0.00306	35214
O	0.80103	0.99967	0.88939	141594
avg / total	0.74797	0.80087	0.71286	176808



Log Loss: 0.48

y=I top features	
Weight?	Feature
+0.351	0:word.ent_iob_:B
+0.306	0:word.pos_:NUM
+0.306	0:word.tag_:CD
+0.285	0:word.like_num
+0.217	0:word.is_digit()
+0.205	0:word.dep_:pobj
+0.195	0:word.dep_:nsubj
+0.195	-1:word.tag_:``
+0.167	-1:word.pos_:VERB
+0.167	1:word.tag_:''
+0.163	0:word.dep_:appos
+0.156	-1:word.is_stop
… 1264 more positive …	
… 615 more negative …	
-0.155	1:word.dep_:compound
-0.160	-1:word.like_num
-0.180	0:word.dep_:punct
-0.186	0:word.pos_:PUNCT
-0.212	0:word.tag_:VBD
-0.214	1 EOS
-0.215	0:word.pos_:VERB
-0.299	0:word.tag_:.
-0.302	-1:word.dep_:pobj
-0.340	0:word.lemma:.
-0.342	0:word.dep_:ROOT
-0.366	0:word.is_stop
-0.449	1:word.ent_iob_:B

Examples

Formed in November 1990 by the equal merger of Sky Television and British Satellite Broadcasting , BSkyB became the UK 's largest digital subscription television company . Following BSkyB 's 2014 acquisition of Sky Italia and a majority 90.04 % interest in Sky Deutschland in November 2014 , its holding company British Sky Broadcasting Group plc changed its name to Sky plc . The United Kingdom operations also changed the company name from British Sky Broadcasting Limited to Sky UK Limited , still trading as Sky .

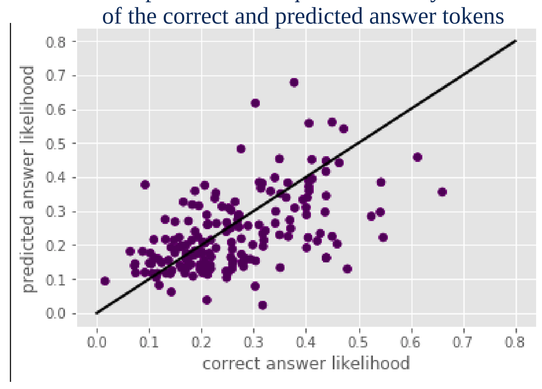
In England , the period of Norman architecture immediately succeeds that of the Anglo - Saxon and precedes the Early Gothic . In southern Italy , the Normans incorporated elements of Islamic , Lombard , and Byzantine building techniques into their own , initiating a unique style known as Norman - Arab architecture within the Kingdom of Sicily .

Evaluation of QA models

- Correct prediction: the predicted answer is a substring of the correct answer (or vice versa)

BERT (Google AI Language)

		True condition	
		answerable	unanswerable
Predicted condition	answerable	4874	989
		179	
	unanswerable	826	4956



BiDAF + Self Attention + ELMo (Allen Institute for AI)

nlnet (Microsoft Research Asia)

Results

- Especially Numbers and Named Entities have a very high propability to be asked about
- This can be caused by a bias in the data set:
  - people were getting paid for creating as many questions as possible
  - it is easier to ask about dates and names
- Modern / state-of-the-art Deep Learning approaches make only a few mistakes (mostly related to unanswerable questions)
  - The likelihood of the answer tokens does not play a major role