Total Recall

Rohan Tilva, Jordan Peykar, Matthew Lee, Bryan Ki, Hannah Cowley

Checkpoint 2 Results

K	Baseline success @k
1	0.1590909090909091
10	0.2840909090909091
100	0.2878787878787879
1000	0.2878787878787879

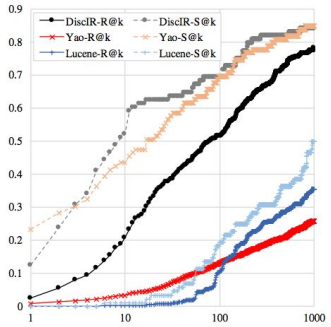


Figure 2: The R@k and S@k curve for different models in the TREC/AQUAINT setting.

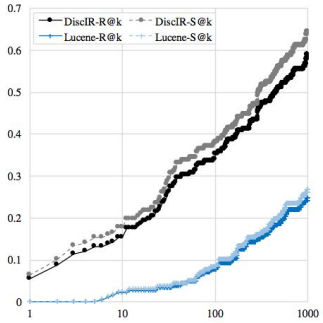


Figure 3: The R@k and S@k curve for different models in the WIKIQA/Wikipedia setting.

```
if query not in used:
    used[query] = 0
    terms = query.split(" ")
    for k_val in k_vals:
        query1 = SearchQuery(type=SearchType.SENTENCES, terms=terms, k=k_val, rawQuery=query)
        results = s.search(query1)
        atK = 0
        totCorrect = 0
        hasAnswerInMatch = False
        for result in results.searchResultItems:
            if atK == k_val:
                break
            else:
                atK += 1
                try:
                    totCorrect += int(answer_labels[result.sentenceId.uuidString])
                    hasAnswerInMatch = True
                except (KeyError):
                    atK -= 1
        if totCorrect >= 1:
            k_val_dict[k_val][0] += 1
        if hasAnswerInMatch:
            k_val_dict[k_val][1] += 1
```

Checkpoint 3: First Iteration

- Classification used to rank potential answers to questions
- Current method:
  - Get TF-IDF scores for terms in questions and answers
  - Calculate cosine similarity
- Optimized C to .0001 using an SVM:
  - Precision: 0.0655966503838
  - Recall: 0.671428571429
  - F1: 0.11951684679

	QA0	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA8	QA9
Cosine Similarity	0.174	0.299	0.321	0.429	0.221	0.184	0.125	0.409	0.475	0.260

Checkpoint 3: Second Iteration

- Implement new feature:
  - Recognize question words
  - Who, What, When Where Why, How
- F1 Score: 0.132530120482
- P: 0.0753424657534
- R: 0.55
- F1 score went up a little (P up, R down)

Other Classification Methods

A Pairwise Classification Method: Preprocessing

- Stopword filtering, all lower case, stemming
  - Q: What is the capital of the United States?
    - what capital united states
  - A: Washington, D.C. is the capital.
    - washington dc capital

  - Q: What is the pope?
    - what pope
  - A: The pope is a cool dude.
    - pope cool dude

Classification: weighted co-occurrence

ANSWER WORDS	QUESTION WORDS: what	capital	united	states	pope
pope	1	0	0	0	1
cool	1	0	0	0	1
dude	1	0	0	0	1
washington	1	1	1	1	0
dc	1	1	1	1	0
capital	1	1	1	1	0

Weighted Bigram co-occurrence

- No stopword filtering
- Q: What is the capital of the United States?
  - [what is] [the capital] [of the] [united states]
  - [what is] [is the] [the capital] [capital of] [of the] [the united] [united states] [states END]
- A: Washington, D.C. is the capital.
  - [washington dc] [is the] [capital END]
  - [washington dc] [dc is] [is the] [the capital]
- *A weighted co-occurrence matrix with bigrams may capture context.*

Weighting by linguistic cues

- Similar to using question words as clues
- Use spaCy to parse out different linguistic features
  - Ie: weight subjects of questions/answers higher
    - If a question and answer share a common subject, more likely to answer the question.
- Positive Example:
  - Q: How big is the **Atlantic Ocean**?
  - A: The **Atlantic Ocean** is big, 41M miles^2
- Negative Example:
  - Q: How big is the **Atlantic Ocean**?
  - A: The **big fish** are in the Atlantic Ocean.

MLP Approach

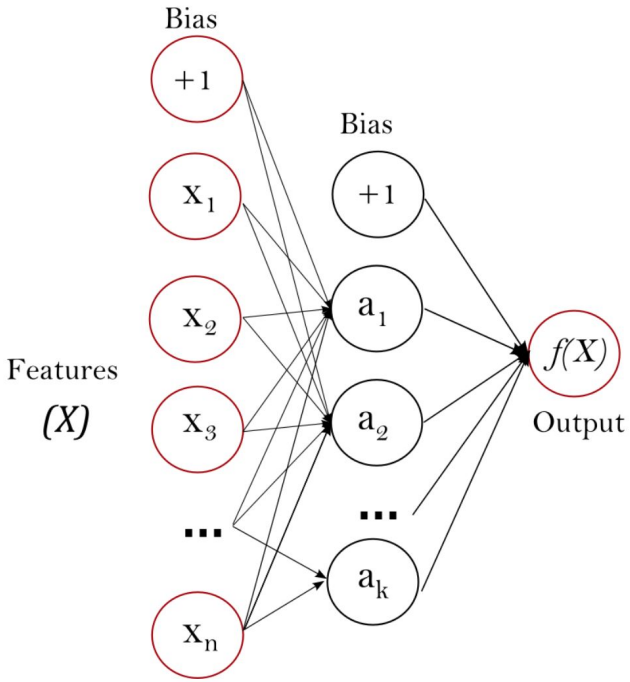


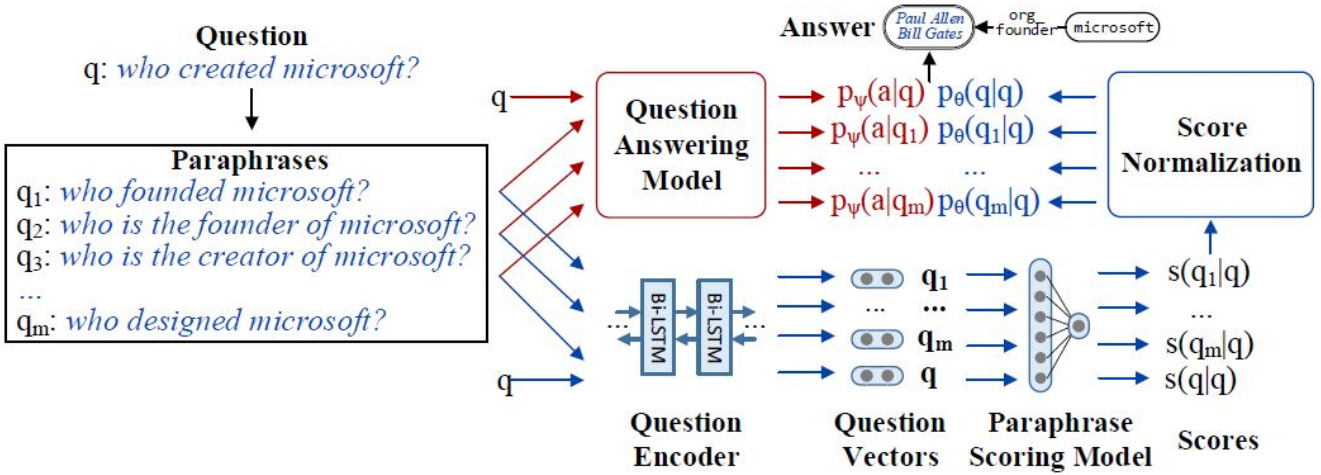
Figure 1 : One hidden layer MLP.

- Feed in a similar feature vector and label vector as SVM
- Add multiple hidden layers to try to sort out non-linearities in the data
- On first pass, saw modest increases in our F1 scores (but still have room to tweak)

Query Expansion

Paraphrasing for Question Answering

- Learning to Paraphrase for Question Answering (2017)
  - (Dong, Mallinson, Reddy, Lapata)



Problem Formulation

Input: what be the zip code of the largest car manufacturer	
what be the zip code of the largest vehicle manufacturer	PPDB
what be the zip code of the largest car producer	PPDB
what be the postal code of the biggest automobile manufacturer	NMT
what be the postcode of the biggest car manufacturer	NMT
what be the largest car manufacturer 's postal code	Rule
zip code of the largest car manufacturer	Rule

$$p\left(a|q\right) = \sum_{q' \in H_q \cup \{q\}} \underbrace{p_\psi\left(a|q'\right)}_{\text{QA Model}} \underbrace{p_\theta\left(q'|q\right)}_{\text{Paraphrase Model}}$$

Paraphrase Generation

- PPDB-based Generation
  - Bilingual pivoting — uses a bilingual parallel corpora to learn paraphrases based on techniques from phrase-based statistical machine translation (SMT, Koehn et al. 2003)
  - car → vehicle manufacturer → producer
  - More local, mostly lexical variation (Pavlick, Ganitkevitch, Van Durme, Callison-Burch)
- NMT-based Generation
  - Neural Machine Translation — maximizes the conditional probability of a correct translation given a source sentence German (Mallinson et al. 2016)
  - Potential to perform major rewrites, generated while considering wider contextual information
- Rule-based Generation
  - WikiAnswers corpus

Source	Target
the average size of __	what be __ average size
__ be locate on which continent	what continent be __ a part of
language speak in __	what be the official language of __
what be the money in __	what currency do __ use

Results

Method	Average F1 (%)	
	GRAPHQ	WEBQ
SEMPRE (Berant et al., 2013)	10.8	35.7
JACANA (Yao and Van Durme, 2014)	5.1	33.0
PARASEMP (Berant and Liang, 2014)	12.8	39.9
SUBGRAPH (Bordes et al., 2014a)	-	40.4
MCCNN (Dong et al., 2015)	-	40.8
YAO15 (Yao, 2015)	-	44.3
AGENDAIL (Berant and Liang, 2015)	-	49.7
STAGG (Yih et al., 2015)	-	48.4 (52.5)
MCNN (Xu et al., 2016)	-	47.0 (53.3)
TYPERERANK (Yavuz et al., 2016)	-	**51.6** (52.6)
BILAYERED (Narayan et al., 2016)	-	47.2
UDEPLAMBDA (Reddy et al., 2017)	**17.6**	49.5
SIMPLEGRAPH (baseline)	15.9	48.5
AVGPARA	16.1	48.8
SEPPARA	18.4	49.6
DATAAUGMENT	16.3	48.7
PARA4QA	**20.4**	**50.7**
−NMT	18.5	49.5
−PPDB	19.5	50.4
−RULE	19.4	49.1

Examples	$p_\theta\,(q'\|q)$
(`music.concert_performance.performance_role`)	
what sort of part do queen play in concert	0.0659
what role do queen play in concert	0.0847
what be the role play by the queen in concert	0.0687
what role do queen play during concert	0.0670
what part do queen play in concert	0.0664
which role do queen play in concert concert	0.0652
(`sports.sports_team_roster.team`)	
what team do shaq play 4	0.2687
what team do shaq play for	0.2783
which team do shaq play with	0.0671
which team do shaq play out	0.0655
which team have you play shaq	0.0650
what team have we play shaq	0.0497

Verb Variations in Question/Answer pair

- Q: “What movies has Harrison Ford acted in?”
- If the answer has variations of the verb “act”, could be correct
  - “Harrison Ford **acts** in Star Wars.”
  - “Ford has **acted** in Star Wars.”
  - “Ford is known for **acting** in Star wars.”
- Different tenses -- stemming

WordNet synonyms

- Building off of suggestion given in checkpoint 3
- Q: “What are synonyms for the verb run?”
- Enter WordNet
  - Look for verb synonyms
  - “Sprint”
  - “Race”
  - “Dash”
- If these synonyms are in the result, higher likelihood of being correct answer

First Word in Question

- First word in question can give lots of information
  - “Who”
  - “Where”
  - “When”
- If “Who”, check for person in answer
  - Capitalized first letters of words (like in “Rohan Tilva”)?
  - Words not recognized by dictionary?
- If “Where”, check for named entity of type location in answer
- If “When”, check for date/numbers in answer (ie. “1948”).

Word Embeddings

- spaCy
- Allows for easy vectorization of words so that words can be compared using similarity vectors rather than or in addition to TF-IDF vectors
- Similarities can also be calculated within context
- Also can do Sentence similarity
- Pitfalls:
  - Not all words are in the vocabulary
  - Hopefully we can catch most of these words with Entity Recognition

Entity Recognition

- Stanford NER (Finkel et. al.) requires labeled data
- spaCy also requires labeled data
- Datasets available (CoNLL 2003)
- Scholarly usage of NER resulted in F1 scores as high as .71 on dev .69 on test

Jim bought 300 shares of Acme Corp. in 2006.

$[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$ in $[\text{2006}]_{\text{Time}}$.

Other Augmentations

Parallelization of computations

- Performance related to speed of computation rather than accuracy
- Can run similarity metrics on multiple threads