Using Semantic and Context Features for Answer Summary Extraction

Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, Mark Sanderson
RMIT University

Why Answer Summaries Matter?

Query: what are some of the possible complications and potential dangers of gastric bypass surgery?

Err… where's the answer?

Bariatric surgery Risks - Mayo Clinic
www.mayoclinic.org/tests-procedures/**bariatric-surgery**/basics/**risks**/prc-20019138
It's important to understand **risks** and results of **gastric bypass** and other types … **gastric bypass** and other **weight-loss surgeries** pose **potential** health **risks**, both … Longer term **risks** and **complications** of **weight-loss surgery** vary depending on ...

Indicative, but not informative

- Satisfying user needs more quickly
- "Good abandonment"
- Beneficial to mobile search

Background

- Passage-based answer extraction:
  - Statistical translation[1]
  - Paid crowdsourcing[2]
  - Query likelihood passage retrieval[3]

- Topical relevance: ineffective for finding answers[3]
- **Summarization** was not considered previously
  - QA used to be “factoid”-based

[1] Soricut and Brill. Automatic Question Answering Using the Web: Beyond the Factoid. Inf. Retr., 9(2). 2006.
[2] Bernstein et al. Direct Answers for Search Queries in the Long Tail. In Proc. of SIGCHI, pages 237{246, 2012.
[3] Keikha et al. Retrieving Passages and Finding Answers. In Proc. of ADCS, pages 81-84, 2014.

Challenges

Locating answer-bearing sentences

1. Vocabulary mismatch
   - Questions are worded differently in the docs

2. Target mismatch
   - Answers are in nearby sentences
   - Discourse might be helpful but too costly

Dataset

WebAP (Keikha et al., 2014)

- 82 queries, with top docs sentence-delimited
- Passage-level annotation in 4 levels:
  - Only relevant documents were involved
  - 80 queries, 1436 docs, and 3298 answers
  - **Perfect** and **Excellent** sents were used (~ 6%)

> Another common complication from gastric bypass is "dumping syndrome.“ The symptoms often include: * Nausea and vomiting * Diarrhea * Bloated feeling * Dizziness * Sweating.

Example of the perfect answer for query “*what are some of the possible complications and potential dangers of gastric bypass surgery*”

Experimental Setup

Table 1: List of features

MK	Exact Match	Binary value indicating the query being of substring in the sentence
	Term Overlap	Fraction of query terms that occur in the sentence
	Synonym Overlap	Fraction of query terms as well as their synonyms that occur in the sentence
	Language Model Score	Log-likelihood of the query generated from the sentence [8]
	Sentence Length	Number of terms in the sentence
	Sentence Location	Relative location of the sentence within the document
Sem	ESA	Cosine similarity between the query and the sentence ESA vectors
	Word2Vec	Average pairwise cosine similarity between any query and sentence word vectors
	TAGME	Jaccard coefficient between the query and the sentence entity sets
Con	X_{before}	Feature X of the sentence immediately before this sentence
	X_{after}	Feature X of the sentence immediately after this sentence

Model trained in a “per-doc” basis. Top 3 sentences as the summary. 5-fold CV.

Baseline: CNN[2] and the original MK approach

[1] Yang et al. Beyond factoid QA: Effective methods for non-factoid answer sentence retrieval. ECIR ‘16.
[2] Severyn and Moschitti. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. SIGIR ‘15.

Results

Method		R-1	R-2	R-SU4	N@3	P@3
CNN[1]		0.550	0.318	0.343	0.196	0.164
MK	MART	0.599	0.365	0.389	0.229	0.183
MK+Sem		0.619	0.396†	0.417†	0.260†	0.212‡
MK+Sem+Con		0.632‡	0.427‡**	0.447‡**	0.300‡**	0.246‡**
MK	Lambda MART	0.586	0.354	0.377	0.231	0.179
MK+Sem		0.619‡	0.426‡	0.446‡	0.280‡	0.226‡
MK+Sem+Con		**0.661‡****	**0.466‡****	**0.484‡****	**0.340‡****	**0.268‡****

†/‡: $p < 0.05/0.01$ wrt MK   */**: $p < 0.05/0.01$ wrt MK+Sem

[1] https://github.com/aseveryn/deep-qa (Severyn and Moschitti, 2015)

Results

Table 4: Top 5 features. Significant decreases of ROUGE-2 scores induced by the feature ablations are indicated by †/‡ (for $p<0.05$ and $p<0.01$)

No	Feature	Category	Decrease in R-2
1	ESA	Semantic	0.043‡ (-9.23%)
2	TAGME	Semantic	0.025 (-5.36%)
3	$\text{Length}_{\text{after}}$	Context	0.018 (-3.86%)
4	$\text{SynOverlap}_{\text{after}}$	Context	0.015 (-3.22%)
5	LM	MK	0.014 (-3.00%)

Table 5: Correlation between Measures

	R-2	R-SU4	N@3	P@3
R-1	0.922‡	0.945‡	0.564‡	0.520‡
R-2	–	0.985‡	0.659‡	0.617‡
R-SU4	–	–	0.644‡	0.599‡
N@3	–	–	–	0.855‡

Conclusion

- CNN model struggles on this task
  - Non-factoid QA is challenging
- Improvements in answer quality is significant
  - Semantics/context info helps
  - Doubly confirmed with ablation analysis
- Moderate correlation between ROUGE score and ranking measures

Thank you