Review:LSTM-Based Deep Learning Model for Nonfactoid Answer Selection

Under review as a conference paper at ICLR 2016

An Weijie

Motivation

- Apply a general deep learning (DL) framework for the answer selection task
- Combining convolutional neural network & Utilize a simple but efficient attention to generate the answer representation according to the question context

Main Challenge

- The major challenge of this task is that the correct answer might not directly share lexical units with the question.
- The answers are sometimes noisy and contain a large amount of unrelated information.

Basic Model: QA-LSTM

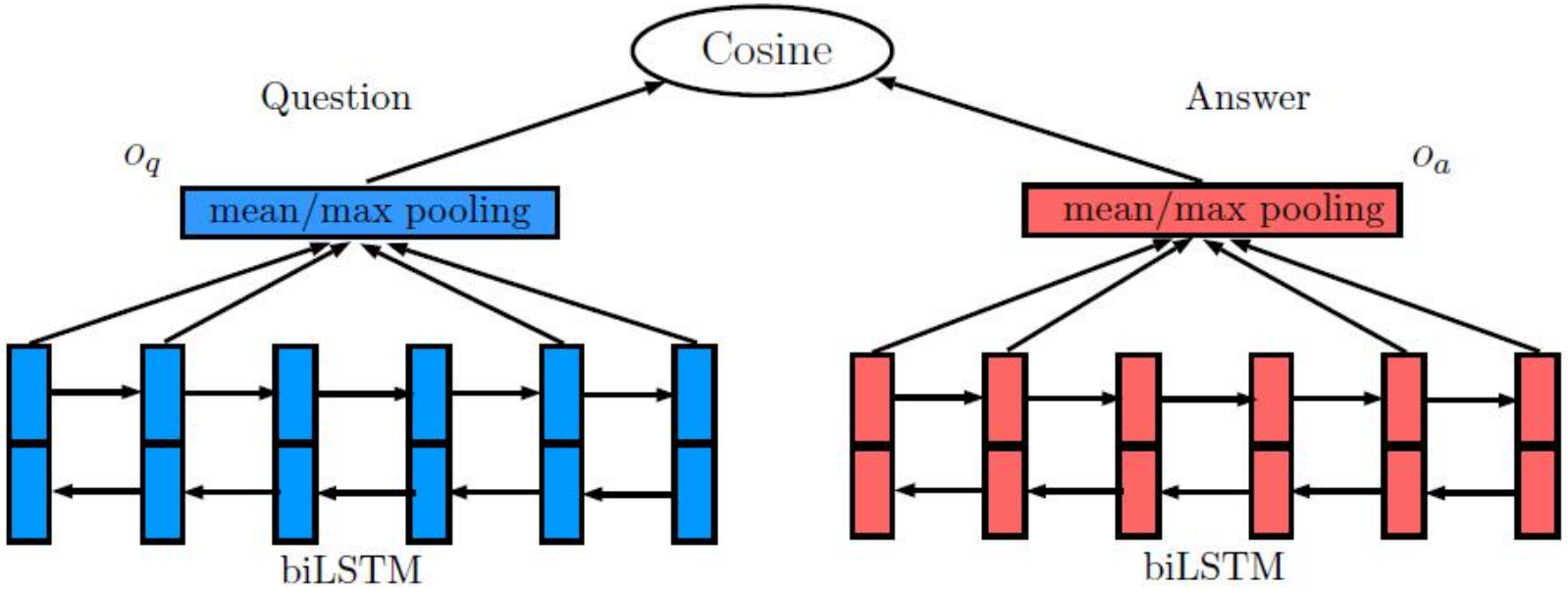

Figure 1: Basic Model: QA-LSTM

QA-LSTM/CNN

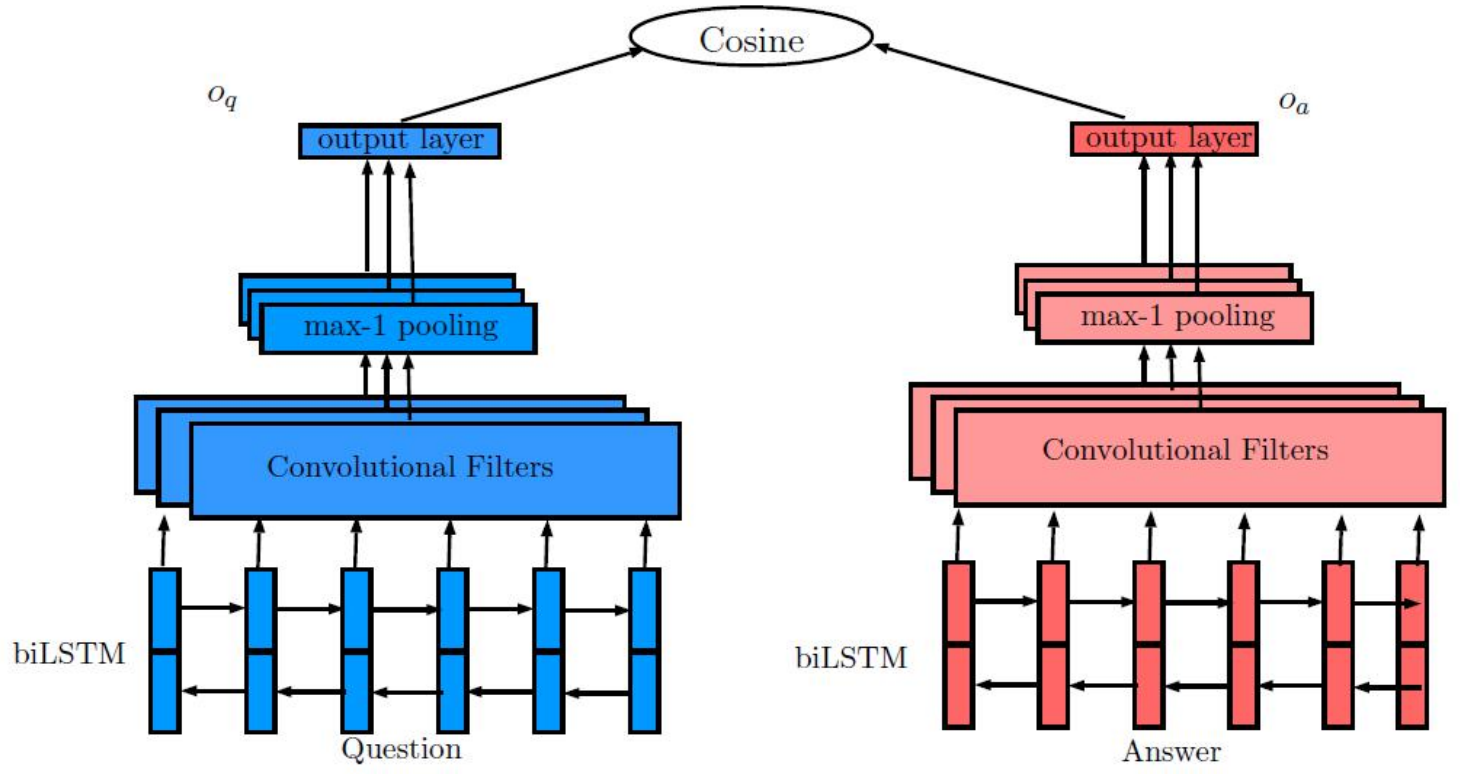

Figure 2: QA-LSTM/CNN

More composite representation of questions and answers.

$$o_F(t) = \tanh\left[\left(\sum_{i=0}^{m-1} \mathbf{h}(t+i)^T \mathbf{F}(i)\right) + b\right]$$

Attention-Based QA-LSTM

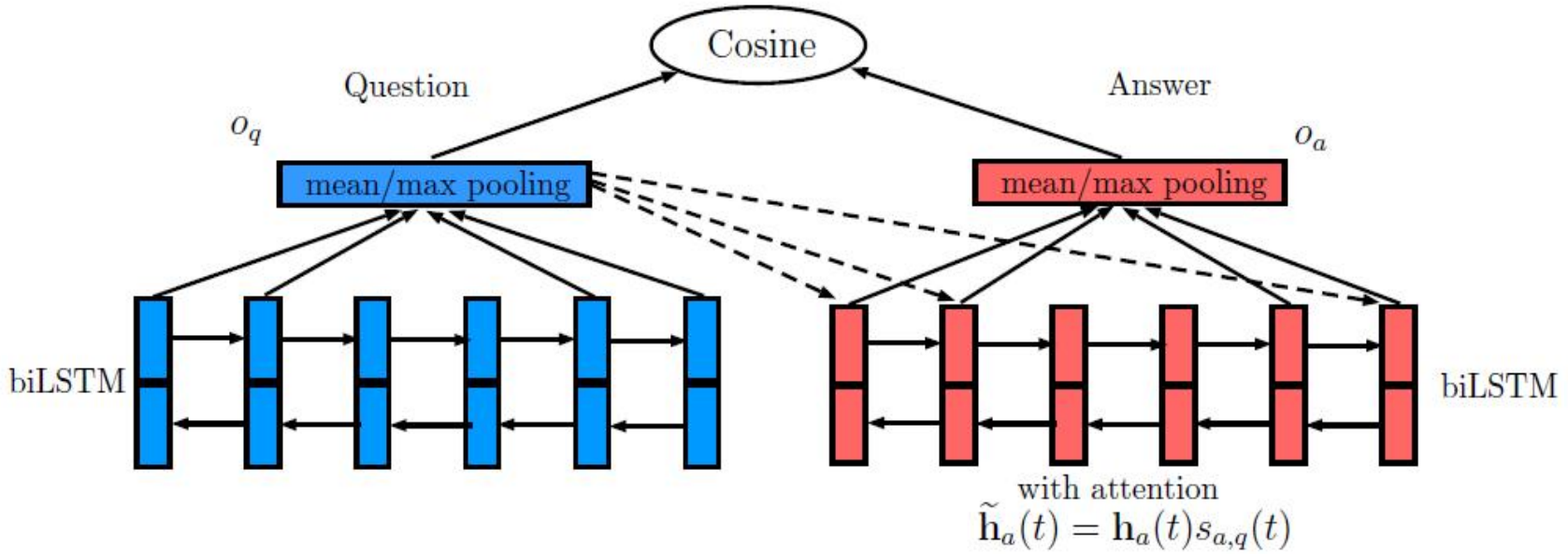

Figure 3: QA-LSTM with attention

Use a simple attention model for the answer vector generation based on questions

$$
\begin{aligned}
\mathbf{m}_{a,q}(t) &= \tanh(\mathbf{W}_{am}\mathbf{h}_a(t) + \mathbf{W}_{qm}\mathbf{o}_q) \\
s_{a,q}(t) &\propto \exp(\mathbf{w}_{ms}^T\mathbf{m}_{a,q}(t)) \\
\widetilde{\mathbf{h}}_a(t) &= \mathbf{h}_a(t)s_{a,q}(t)
\end{aligned}
$$

Experiments

- InsuranceQA Experiment

  (https://github.com/shuzi/insuranceQA.git)

- TrecQA Experiment

  (http://cs.jhu.edu/~xuchen/packages/jacana-qa-naacl2013-data-results.tar.bz2)

InsuranceQA

	Validation	Test1	Test2
A. Bag-of-word	31.9	32.1	32.2
B. Metzler-Bendersky IR model	52.7	55.1	50.8
C. Architecture-II in (Feng et al., 2015)	61.8	62.8	59.2
D. Architecture-II with GESD	**65.4**	**65.3**	**61.0**

Table 2: Baseline results of InsuranceQA

	Model	Validation	Test1	Test2
A	QA-LSTM basic-model(head/tail)	54.0	53.1	51.2
B	QA-LSTM basic-model(avg pooling)	58.5	58.2	54.0
C	QA-LSTM basic-model(max pooling)	64.3	63.1	58.0
D	QA-LSTM/CNN(fcount=1000)	65.5	65.9	62.3
E	QA-LSTM/CNN(fcount=2000)	64.8	66.8	62.6
F	QA-LSTM/CNN(fcount=4000)	66.2	64.6	62.2
G	QA-LSTM with attention (max pooling)	66.5	63.7	60.3
H	QA-LSTM with attention (avg pooling)	**68.4**	**68.1**	62.2
I	QA-LSTM/CNN (fcount=4000) with attention	67.2	65.7	**63.3**

TrecQA

Models	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.6029	0.6852
Yao et al. (2013)	0.6307	0.7477
Severyn & Moschitti (2013)	0.6781	0.7358
Yih et al. (2013)-BDT	0.6940	0.7894
Yih et al. (2013)-LCLR	0.7092	0.7700
Wang & Nyberg (2015)	**0.7134**	0.7913
Architecture-II (Feng et al., 2015)	0.7106	**0.7998**

Table 4: Test results of baselines on TREC-QA

	Models	MAP	MRR
A	QA-LSTM (avg-pool)	68.19	76.52
B	QA-LSTM with attention	68.96	78.49
C	QA-LSTM/CNN	70.61	81.04
D	QA-LSTM/CNN with attention	**71.11**	**83.22**
E	QA-LSTM/CNN with attention (LSTM hiddenvector=500)	**72.79**	**82.40**

Table 5: Test results of the proposed models on TREC-QA

What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA

Mengqiu Wang and **Noah A. Smith** and **Teruko Mitamura**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{mengqiu,nasmith,teruko}@cs.cmu.edu

Models	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.6029	0.6852
Yao et al. (2013)	0.6307	0.7477
Severyn & Moschitti (2013)	0.6781	0.7358
Yih et al. (2013)-BDT	0.6940	0.7894
Yih et al. (2013)-LCLR	0.7092	0.7700
Wang & Nyberg (2015)	**0.7134**	0.7913
Architecture-II (Feng et al., 2015)	0.7106	**0.7998**

Table 4: Test results of baselines on TREC-QA

training dataset	model	development set MAP	development set MRR	test set MAP	test set MRR
100 manually-judged	TreeMatch	0.4074	0.4458	0.3814	0.4462
	+WN	0.4328	0.4961	0.4189	0.4939
	Cui et al.	0.4715	0.6059	0.4350	0.5569
	+WN	0.5311	0.6162	0.4271	0.5259
	Jeopardy (base only)	0.5189	0.5788	0.4828	0.5571
	Jeopardy	**0.6812**	**0.7636**	**0.6029**	**0.6852**
+2,293 noisy	Cui et al.	0.2165	0.3690	0.2833	0.4248
	+WN	0.4333	0.5363	0.3811	0.4964
	Jeopardy (base only)	0.5174	0.5570	0.4922	0.5732
	Jeopardy	0.6683	0.7443	0.5655	0.6687

Tree Edit Models for Recognizing Textual Entailments, Paraphrases, and Answers to Questions

Michael Heilman Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{mheilman,nasmith}@cs.cmu.edu

Models	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.6029	0.6852
Yao et al. (2013)	0.6307	0.7477
Severyn & Moschitti (2013)	0.6781	0.7358
Yih et al. (2013)-BDT	0.6940	0.7894
Yih et al. (2013)-LCLR	0.7092	0.7700
Wang & Nyberg (2015)	**0.7134**	0.7913
Architecture-II (Feng et al., 2015)	0.7106	**0.7998**

Table 4: Test results of baselines on TREC-QA

System	MAP	MRR
Punyakanok et al., 2004	0.3814	0.4462
+WN	0.4189	0.4939
Cui et al., 2005	0.4350	0.5569
+WN	0.4271	0.5259
Wang et al., 2007	0.4828	0.5571
+WN	0.6029	0.6852
Tree Edit Model	**0.6091**	**0.6917**

Table 5: Results for the task of answer selection for question answering. +WN denotes use of WordNet features.

Probabilistic Tree-Edit Models with Structured Latent Variables for Textual Entailment and Question Answering

Mengqiu Wang
Computer Science Department
Stanford University
mengqiu@cs.stanford.edu

Christopher D. Manning
Computer Science Department
Stanford University
manning@cs.stanford.edu

Models	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.6029	0.6852
Yao et al. (2013)	0.6307	0.7477
Severyn & Moschitti (2013)	0.6781	0.7358
Yih et al. (2013)-BDT	0.6940	0.7894
Yih et al. (2013)-LCLR	0.7092	0.7700
Wang & Nyberg (2015)	**0.7134**	0.7913
Architecture-II (Feng et al., 2015)	0.7106	**0.7998**

Table 4: Test results of baselines on TREC-QA

System	MAP	MRR
Punyakanok et al., 2004	0.4189	0.4939
Cui et al., 2005	0.4350	0.5569
Wang et al., 2007	0.6029	0.6852
H&S, 2010	**0.6091**	0.6917
Tree-edit CRF	0.5951	**0.6951**

Table 3: Results on QA task reported in Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR).

Answer Extraction as Sequence Tagging with Tree Edit Distance

Xuchen Yao and **Benjamin Van Durme**
Johns Hopkins University
Baltimore, MD, USA

Models	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.6029	0.6852
Yao et al. (2013)	0.6307	0.7477
Severyn & Moschitti (2013)	0.6781	0.7358
Yih et al. (2013)-BDT	0.6940	0.7894
Yih et al. (2013)-LCLR	0.7092	0.7700
Wang & Nyberg (2015)	**0.7134**	0.7913
Architecture-II (Feng et al., 2015)	0.7106	**0.7998**

Table 4: Test results of baselines on TREC-QA

System	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman and Smith (2010)	0.6091	0.6917
Wang and Manning (2010)	0.5951	0.6951
this paper (48 features)	0.6319	0.7270
+WNsearch	**0.6371**	0.7301
+WNfeature (11 more feat.)	0.6307	**0.7477**

Table 3: Results on the QA Sentence Ranking task.

Automatic Feature Engineering for Answer Selection and Extraction

Aliaksei Severyn
DISI, University of Trento
38123 Povo (TN), Italy
severyn@disi.unitn.it

Alessandro Moschitti
Qatar Computing Research Institue
5825 Doha, Qatar
amoschitti@qf.org.qa

Models	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.6029	0.6852
Yao et al. (2013)	0.6307	0.7477
Severyn & Moschitti (2013)	0.6781	0.7358
Yih et al. (2013)-BDT	0.6940	0.7894
Yih et al. (2013)-LCLR	0.7092	0.7700
Wang & Nyberg (2015)	**0.7134**	0.7913
Architecture-II (Feng et al., 2015)	0.7106	**0.7998**

Table 4: Test results of baselines on TREC-QA

Table 4: Answer sentence reranking on TREC 13.

System	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.5951	0.6951
Yao et al. (2013)	0.6319	0.7270
+ WN	0.6371	0.7301
shallow tree (S&M, 2012)	0.6485	0.7244
+ semantic tagging	**0.6781**	**0.7358**

Question Answering Using Enhanced Lexical Semantic Models

Wen-tau Yih **Ming-Wei Chang** **Christopher Meek** **Andrzej Pastusiak**
Microsoft Research
Redmond, WA 98052, USA
{scottyih,minchang,meek,andrzejp}@microsoft.com

Models	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.6029	0.6852
Yao et al. (2013)	0.6307	0.7477
Severyn & Moschitti (2013)	0.6781	0.7358
Yih et al. (2013)-BDT	0.6940	0.7894
Yih et al. (2013)-LCLR	0.7092	0.7700
Wang & Nyberg (2015)	**0.7134**	0.7913
Architecture-II (Feng et al., 2015)	0.7106	**0.7998**

Table 4: Test results of baselines on TREC-QA

	LR		BDT		LCLR	
Feature set	MAP	MRR	MAP	MRR	MAP	MRR
1: I	0.6531	0.7071	0.6323	0.6898	0.6629	0.7279
2: I+L	0.6744	0.7223	0.6496	0.6923	0.6815	0.7270
3: I+L+WN	0.7039	0.7705	0.6798	0.7450	0.7316	0.7921
4: I+L+WN+LS	0.7339	0.8107	0.7523	0.8455	0.7626	0.8231
5: All	0.7374	0.8171	0.7495	0.8450	0.7648	0.8255

A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering

Di Wang and **Eric Nyberg**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{diwang,ehn}@cs.cmu.edu

Models	MAP	MRR
Wang et al. (2007)	0.6029	0.6852
Heilman & Smith (2010)	0.6091	0.6917
Wang & Manning (2010)	0.6029	0.6852
Yao et al. (2013)	0.6307	0.7477
Severyn & Moschitti (2013)	0.6781	0.7358
Yih et al. (2013)-BDT	0.6940	0.7894
Yih et al. (2013)-LCLR	0.7092	0.7700
Wang & Nyberg (2015)	**0.7134**	0.7913
Architecture-II (Feng et al., 2015)	0.7106	**0.7998**

Table 4: Test results of baselines on TREC-QA

Features	MAP	MRR
BM25	0.6370	0.7076
Single-Layer LSTM	0.5302	0.5956
Single-Layer BLSTM	0.5636	0.6304
Three-Layer BLSTM	0.5928	0.6721
Three-Layer BLSTM + BM25	**0.7134**	**0.7913**

TrecQA 8-13

http://cs.jhu.edu/~xuchen/packages/jacana-qa-naacl2013-data-results.tar.bz2

```
<QApairs id='1'>
<question>
Who is the author of the book , `` The Iron Lady : A Biography of Margaret Thatcher
'' ?
NNP VBZ DT NN IN DT NN , `` DT JJ NN : DT NN IN NNP NNP '' .
SUB ROOT NMOD PRD NMOD NMOD PMOD P P NMOD NMOD NMOD P NMOD NMOD NMOD
NMOD PMOD P P
2 0 4 2 4 7 5 15 15 12 12 15 15 15 7 15 18 16 18 2
- - - PER_DESC-B - - - - - WORK_OF_ART-B WORK_OF_ART-I WORK_OF_ART-I WORK_OF_ART-I
WORK_OF_ART-I WORK_OF_ART-I - PERSON-B PERSON-I - -
</question>
<positive>
the IRON LADY ; A Biography of Margaret Thatcher by Hugo Young -LRB- Farrar ,
Straus & Giroux -RRB-
DT NNP NNP : DT NN IN NNP NNP IN NNP NNP -LRB- NNP , NNP CC NNP -RRB-
NMOD NMOD NMOD P NMOD ROOT NMOD NMOD PMOD NMOD NMOD PMOD P NMOD P
NMOD NMOD PMOD P
3 3 6 6 6 0 6 9 7 6 12 10 18 18 18 18 18 10 18
- ORGANIZATION-B ORGANIZATION-I ORGANIZATION-I ORGANIZATION-I ORGANIZATION-I - PERSON-B PERSON-I
- PERSON-B PERSON-I - ORGANIZATION-B ORGANIZATION-I ORGANIZATION-I ORGANIZATION-I
ORGANIZATION-I -
Hugo Young
11 12
</positive>
```

1162 training questions, 65 development questions and 68 test questions.

InsuranceQA

https://github.com/shuzi/insuranceQA.git

- For all train/valid/test files, format is same, with various answer pool size:
  - `<Domain><TAB><QUESTION><TAB><Groundtruth><TAB><Pool>`
- For InsuranceQA.question.anslabel.*:
  - `<Domain><TAB><QUESTION><TAB><Groundtruth>`
- For InsuranceQA.label2answer.*
  - `<Answer Label><TAB><Answer Text>`
- For vocabulary file:
  - `<word index><TAB><original word>`

Corpus Statistics

	Question	Answer	Question Running Words
Train	12,889	21,325	107,889
Valid	2,000	3354	16,931
Test	2,000	3308	16,815