Leveraging External Knowledge in Natural Language Understanding Systems

Jackie Chi Kit Cheung

Mila / McGill University

jcheung@cs.mcgill.ca

August 31st, 2018

Understanding by Reading

Entities

Events

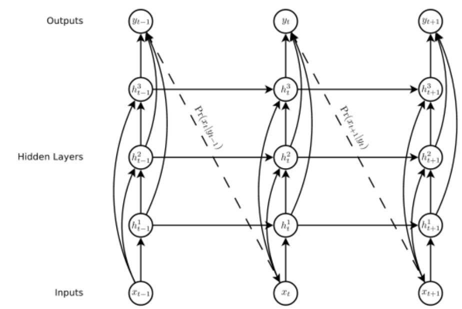
Relations

Desires

Beliefs

Attitudes

Intentions



Applications:

Text classification

Sentiment analysis

Automatic summarization

Question answering

Machine translation

Traditional Supervised Learning

1. Gather training data
2. Annotate, annotate, annotate
3. Train supervised classifier

Problems:

- Rare or new words and entities
- Learned cues often don't transfer between tasks

Transferring Between Tasks

Supervised learning tends to pick up on cues that don't transfer between tasks.

- e.g., pick up on words that happen to frequently co-occur with target class

What's wrong with this approach?

- Using surface cues rather than semantic understanding
- "Tricks" to pass the Turing test

Predicting Dialogue Success

Created a dataset from stackoverflow.com for predicting success in goal-driven human-human dialogues

User A :	I accidentally closed the Stack Trace window in the Visual Studio 2008 debugger. How do I redisplay this window?
User B :	While debugging: Debug\Windows\Call stack
User A :	Thanks, I don't know how I overlooked it.

Noseworthy, Cheung, and Pineau, *SIGDIAL 2017*

What Do Current Methods Do?

Long Short-Term Memory Networks (Hochreiter and Schmidhuber, 1997)

Information Given	Success F1	Failure F1
Full conversation thread	89	73
Only the last comment	86	68
Without last comment	83	38

- Suggests LSTMs mostly capture discourse cues
- Task-specific supervised learning focuses on useful cues for this task only!
- **Much harder**: understanding whether information need in question was satisfied

Case Study: Turing Test – Solved?!

Computer AI passes Turing test in 'world first'

9 June 2014 | Technology

Share

Eugene Goostman simulates a 13-year-old Ukrainian boy

A computer program called Eugene Goostman, which simulates a 13-year-old Ukrainian boy, is said to have passed the Turing test at an event organised by the University of Reading.

No Real World or Conversation Understanding

Do you have an iPhone?

If I say that I haven't an iphone - will it satisfy your inquisitive mind? <u>And I forgot to ask you where you are from...</u>

You asked me where I was from already.

So what that you were from already By the way, what's your occupation? I mean - could you tell me about your work?

This is now unraveling at a rapid clip.

You're not the first one who told me that.

Nice save. You're back in the game.

Possibly you are right. You aren't the first one who tell me that. Oooh. Anything else?

I think we're done here. Congrats on the thing.

Your opinion is very interesting... Keep going. <u>And I forgot to ask you where you are from...</u>

http://time.com/2847900/eugene-goostman-turing-test/

External Knowledge

Use external knowledge to help us overcome these barriers!

Already widespread:

- Pre-trained word embeddings (word2vec, GloVe)
- Pre-trained language models (Peters et al., 2018; Howard and Ruder, 2018; Radford et al., 2018)

What about more targeted forms of external knowledge?

This Talk

Rare entity prediction

- Target rare or unknown entities
- *External knowledge*: short description of entity

EMNLP 2017

Commonsense reasoning

- Reason about plausibility in the world
- *External knowledge*: entire indexed web

EMNLP 2018

Reading Comprehension

- Read a text
- Understand it
- Answer questions

Several recent datasets:

- Daily Mail/CNN (Hermann et al., 2015)
- SQuAD (Rajpurkar et al., 2016)
- bAbI (Weston et al., 2015)
  - Children's Books (Hill et al., 2015)

What Do Current Tasks Test?

Reasoning *within* the provided passage

- e.g., sense disambiguation, paraphrase recognition

Explicitly try to factor out world knowledge

- Entity anonymization in Daily Mail/CNN:

*"the **ent381** producer allegedly struck by **ent212** will not press charges against the " **ent153** " host , his lawyer said friday . **ent212** , who hosted one of the most - watched television shows in the world , was dropped by the **ent381** wednesday after an internal investigation"*

External Knowledge in Reading

World knowledge and expectations important in human reading (Barrett and Nyhof, 2001)

- Information retention
- Interestingness and importance

AI tradition of organizing information around world knowledge and stereotypical situations

- Scripts (Schank and Abelson, 1977)
- FrameNet (Baker et al., 1998)

Wikilinks Rare Entity Prediction

Wikilinks (Singh et al., 2012)

- Dataset for coreference resolution
- Web corpus where spans are annotated with links to Wikipedia pages
- We enhance this with definitions from Freebase/Wikipedia

Long, Bengio, Lowe, Cheung, Precup
EMNLP 2017

Sample

William Blake, who lived from 1757 to 1827, was admired by a small grou
general recognition as either a poet or painter. Yet today his poems and t
the British psyche in a way that few others can match. *Jerusalem* stirs

William Blake

From Wikipedia, the free encyclopedia

For other people named William Blake, see William Blake (disambiguation).

William Blake (28 November 1757 – 12 August 1827) was an English poet, painter, and printmaker. Largely unrecognised during his lifetime, Blake is now considered a seminal figure in the history of the poetry and visual arts of the Romantic Age. His so-called prophetic works were said by 20th century critic Northrop Frye to form "what is in proportion to its merits the least read body of poetry in the

William Blake

Task Setup

A **plausibility cloze** task:

- Predict which entity from a document fits into a blank, given entity definitions

Context
[...] ______, who lived from 1757 to 1827, was admired by a small group of intellectuals and artists in his day, but never gained general recognition as either a poet or painter. [...]
Candidate Entities
Peter Ackroyd: Peter Ackroyd is an English biographer, novelist and critic with a particular interest in the history and culture of London. [...]
William Blake: William Blake was an English poet, painter, and printmaker. [...]
Emanuel Swedenborg: Emanuel Swedenborg was a Swedish scientist, philosopher, theologian, revelator, and mystic. [...]

Drawn from the same original document

Corpus Characteristics

# Documents	269,469
Avg # entities per doc	3.35
Avg # entity mentions per doc	3.69
# Unique entities	245,116
freq $\leq$ 5	207,435 (84.6%)
freq $\leq$ 10	227,481 (92.8%)
freq $\leq$ 20	238,025 (97.1%)

- Number of entities on par with number of documents!

Double Encoder Model

Given sample i, (i.e., a blank and its context), model calculates

$$P(e|\mathrm{C}_i, \mathrm{L}_e)$$

e Entity

C_i Document context (sentence with blank)

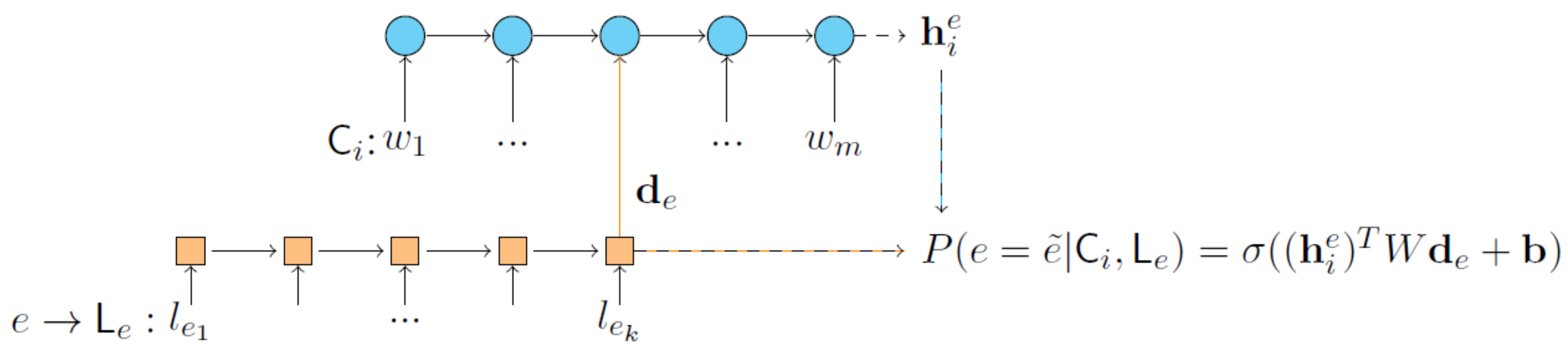
L_e Entity definition (first sentence in Freebase page)

for each candidate entity $\tilde{e}$.

Select entity that maximizes above probability

Double Encoder (DoubEnc)

LSTM encoders for each of C_i and L_e



$\mathbf{d}_e$ and $\mathbf{h}_i^e$ are the encodings of L_e and C_i resp.

W and $\mathbf{b}$ are additional learned parameters

Structure similar to (Bahdanau et al., 2017)

Encoding More Context

Exploit longer-range context: the previous sentences which contain blanks

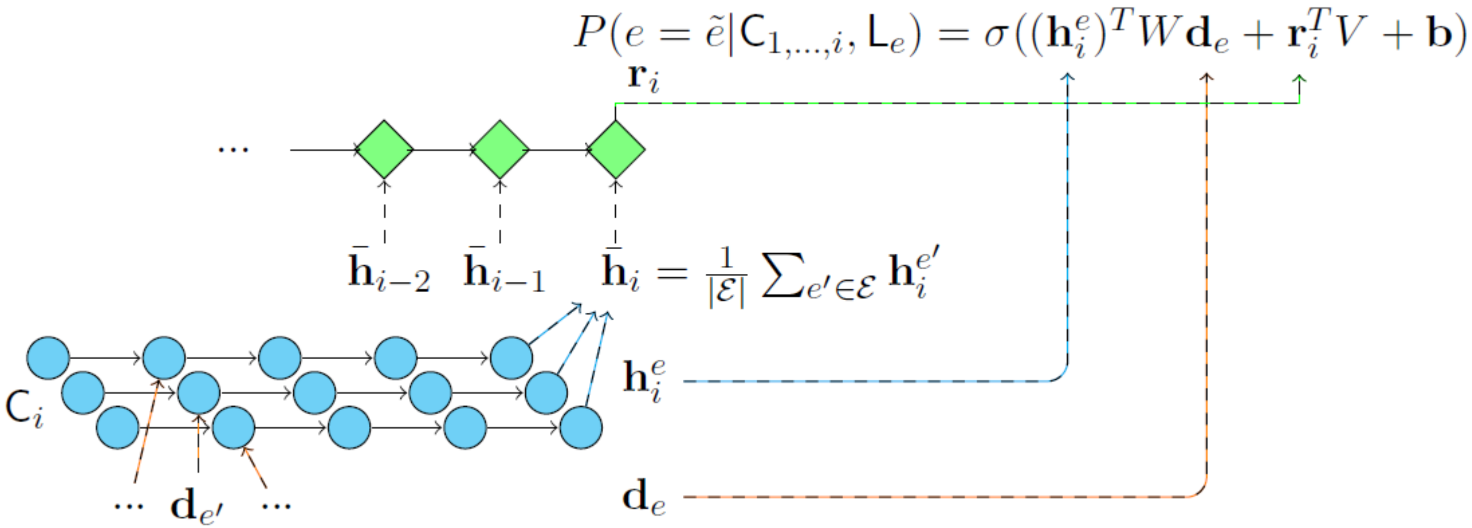
$$P(e|\mathrm{C}_{1\ldots i}, \mathrm{L}_e)$$

For each timestep up to now, compute:

$$\overline{\mathbf{h}_t} = \frac{1}{|\varepsilon|}\sum_{e' \in \varepsilon} \mathbf{h}_i^{e'}$$

Feed these $\overline{\mathbf{h}_t}$ into another LSTM encoder (temporal network)

Hierarchical Encoder (HierEnc)



$\mathbf{r}_i$ is the last hidden layer of the temporal network

V are additional learned parameters

Experiments

Rare entity prediction on our dataset

(80% train/10% dev/10% test)

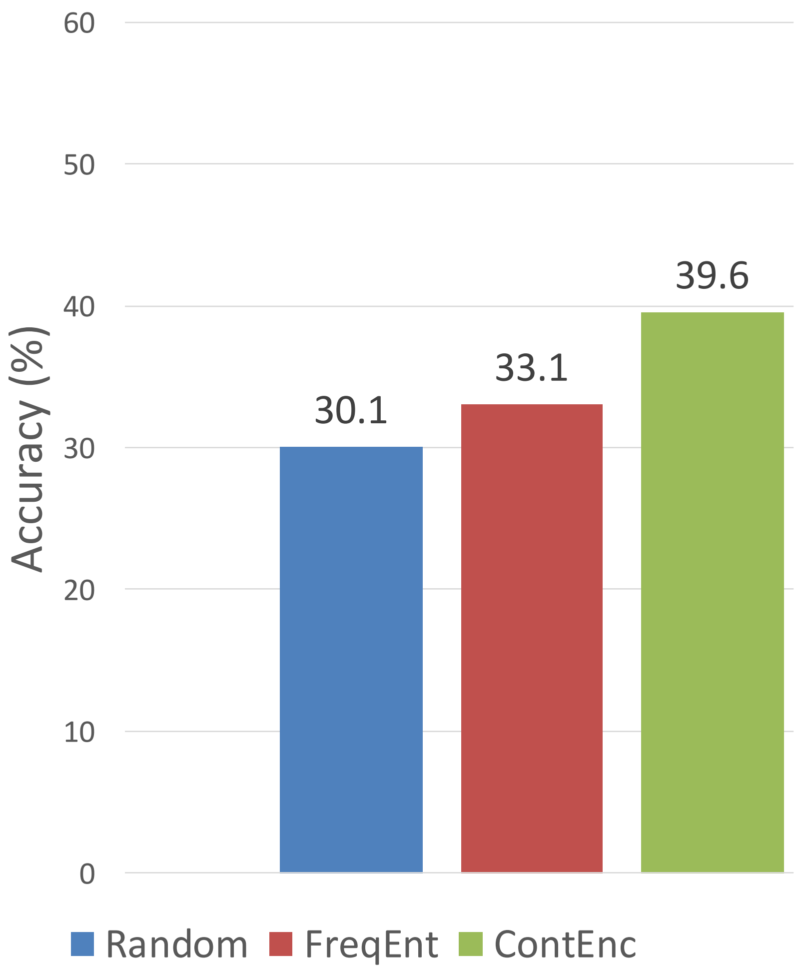
Settings:

- Context window is sentence with blank
- First sentence used in definition
- Binary cross-entropy loss
- Sizes of hidden layers: 300 for context and definition, 200 for temporal network
- SGD training with Adam (Kingma and Ba, 2014)

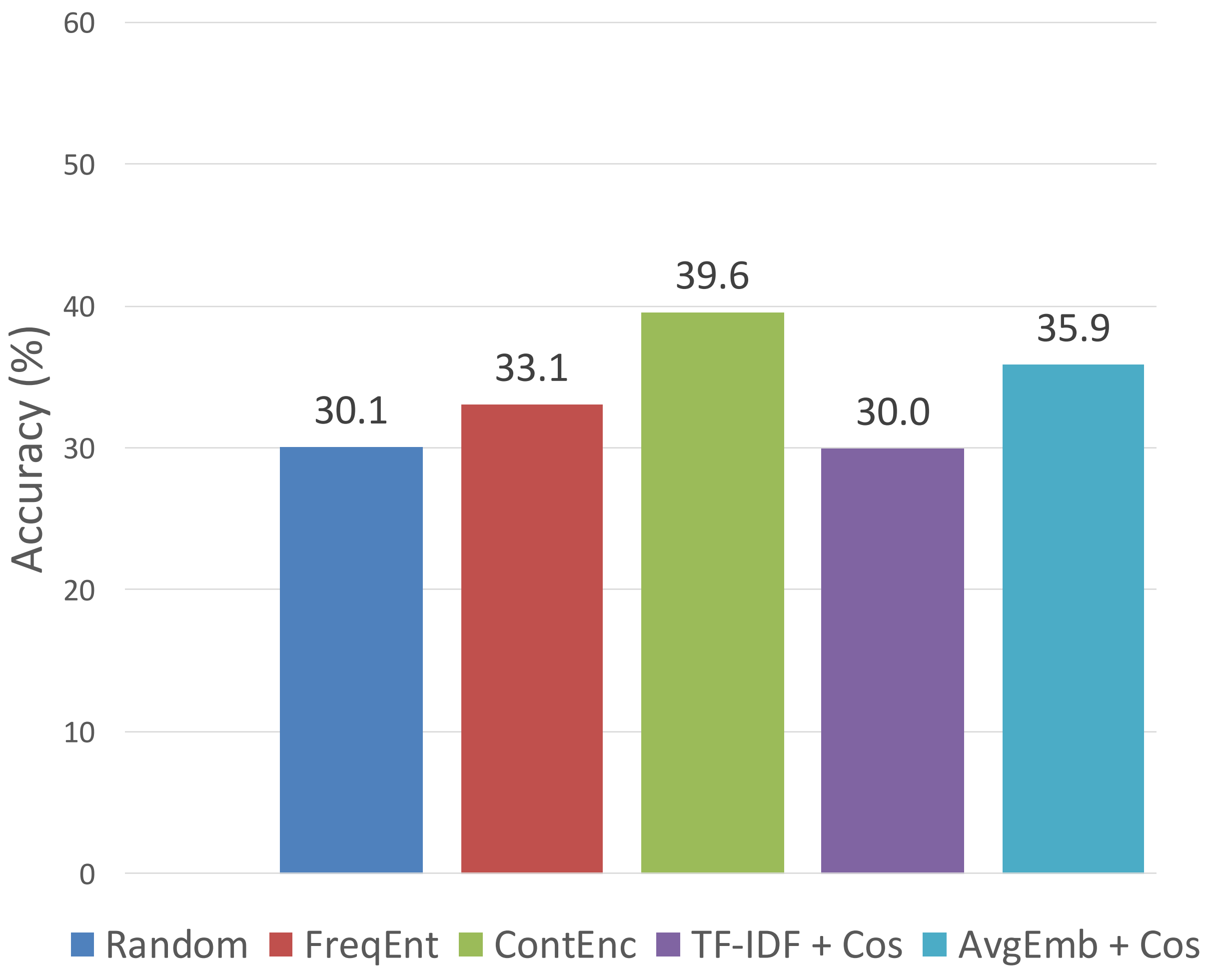
Baselines

Random	Randomly predict an entity
FreqEnt	Select most frequent entity in document
ContEnc	Entities are treated as just another word in an LSTM language model
TF-IDF + Cos	IDF-weighted cosine similarity between definition and context
AvgEmb + Cos	Cosine similarity between average GloVe embeddings of definition and context

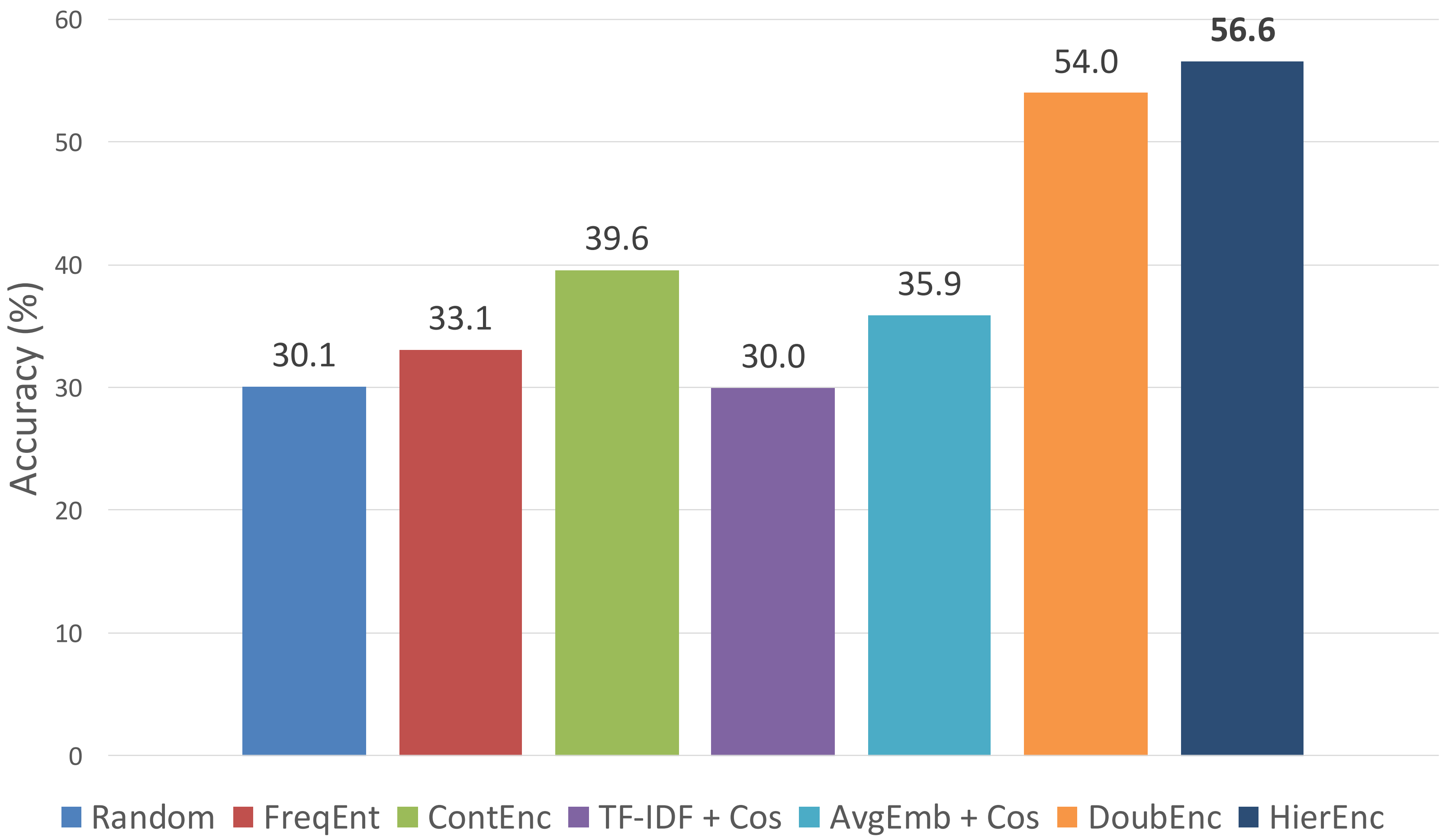
Rare Entity Prediction Results



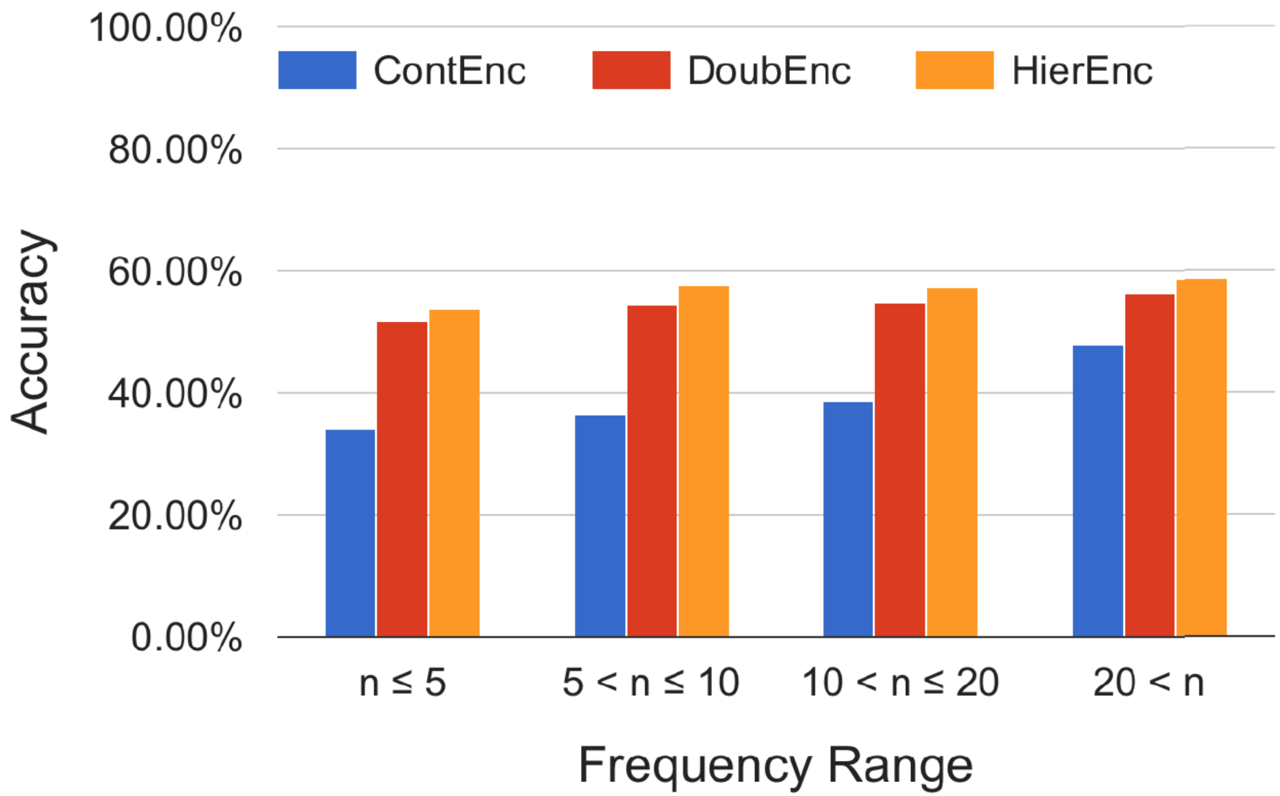
Rare Entity Prediction Results



Rare Entity Prediction Results



Analysis: Entity Frequency



Greater improvement in rarer entities!

Sample Prediction

Context & Prediction

[...] We heard from Audrey Bomse, who is with the Free Gaza movement. She was in ______, Cyprus. [...]

CONTENC: Istanbul   HIERENC: Larnaca

Candidate Set

Istanbul: Istanbul is the most populous city in Turkey, and the country's economic, cultural, and historical center.

Larnaca: Larnaca is a city on the southern coast of Cyprus and the capital of eponymous district.

Ben Macintyre: Ben Macintyre is a British author, historian, reviewer and columnist writing for The Times newspaper.

(Other candidate entities......)

Correct answer: Larnaca

Dataset frequencies:

- Istanbul (86); Larnaca (2, **0** in training)

Common Sense Reasoning

Can we evaluate on a task that is robust to "cheap tricks"? (Levesque, 2013)

- i.e., requires more than word counting

Common sense reasoning data sets which are supposed to be immune to such tricks:

Winograd Schema Challenge

Choice of Plausible Alternatives

Emami, Trischler, Suleman, de la Cruz, Cheung
EMNLP 2018

Winograd Schema Challenge

*The town councilmen refused the protestors a permit because <u>they</u> **feared/advocated** violence.*

- **Hard** coreference resolution questions that require world knowledge
- Answer changes with just a small change in wording!

Winograd Example

The town councilmen refused the protestors a permit because they feared violence.

The town councilmen refused the protestors a permit because they advocated violence.

Because both options appear in dataset, cannot make use of usual syntactic and grammatical cues.

- Pronoun and all potential antecedents agree in grammatical number and gender
- Cannot use simple lexical features, recency, syntactic position, etc. which are usually very useful in coreference resolution (Durrett and Klein, 2013)

Choice of Plausible Alternatives (COPA)

Decide which alternative is more likely:

The climbers reached the peak of the mountain.
What happened as a result?

- *They encountered an avalanche.*
- ***They congratulated each other.***

Dataset is controlled so that alternatives have words that are related to the context.

Entire Web as External Corpus

Rare entity prediction: single sentence as external knowledge for entity being modelled

Now: we need to model the *entire* situation.

Intuition: search for similar situations across the entire indexed web using a search engine!

Pipeline of Our Approach

1. Parse original context
2. Extract search terms from question and search the web for related results
3. Filter search results
4. Reason with search results to make final decision

Schema for WSC Questions

General format of questions:

E1 PredC E2 + P PredQ

Candidate antecedents Context predicate Discourse connective

Target pronoun Query predicate

The trophy didn't fit into the brown suitcase

because it was too {large/small}.

Extracting Search Term

+TermC +TermQ –“Winograd” –E1

TermC and TermQ constructed from PredC and PredQ, with additional modifiers

- TermC ∈ {“doesn’t fit into”, “brown”, “fit”}
- TermQ ∈ {“large”, “is too large”}

Other terms to ensure we don’t just search for the answer on the Winograd website.

Query Augmentation and Filtering

Use WordNet (Kilgarriff 2000) to expand query set with synonyms

- TermC ∈ {“doesn’t fit into”, “brown”, “fit”, **“accommodate”**}
- TermQ ∈ {“large”, “is too large”, **“big”**}

Filtering to remove words in the query terms that are dissimilar to other words (e.g., "brown")

Extracting from Search Results

Search results must fit following form:

E1’ PredC’ E2’ + E3’ PredQ’

Minor variations in word order acceptable

We call these **evidence sentences.**

Run coreference resolution module:

If E3’ corefers with E1’ → **evidence-agent**

If E3’ corefers with E2’ → **evidence-patient**

Pick E1 or E2 depending on which one has more support

Examples

“doesn’t fit into” +“is too large”

E1’ and E3’ corefer (**evidence-agent**)

It is legal to reduce a case that is too large so that it will fit into small claims court. ... A case that doesn't fit into small claims court.

“doesn’t fit into” + “is too small”

E2’ and E3’ corefer (**evidence-patient**)

… who refused to own up to her actual dress size doesn't fit into her dress. .... I'm in a similar situation, my MOH's dress is too small but one of my …

Experiments

Dataset: 273 Winograd Schema questions (135 pairs + 1 triple), 500 COPA questions

Google + Bing as search engine; Stanford CoreNLP to parse question into schema, get coreference resolution results

Evaluation Measures:

Precision:	#Correct / #Answered
Recall:	#Correct / #Dataset
F1:	2 * P * R / (P + R)
Accuracy	(for COPA)

Winograd Schema Challenge Results

Method	Pr	R	F1
Automatic Query Generation	0.56	0.28	0.38
Automatic Query Generation + Synonyms	0.57	0.42	0.48
Automatic Query Generation + Synonyms + Filtering	0.60	**0.44**	**0.51**
Sharma et al., 2015	**0.92**	0.18	0.30

Recently, Trinh and Le propose a language model-based approach trained on 14 different corpora. Best performance of 63.7% accuracy, and they use IR-based methods to construct the training corpora.

COPA Results

Method	Accuracy (%)
Goodwin et al., 2012	63.4
Gordon et al., 2011	65.4
Automatic Query Generation	*66.2*
Luo et al., 2016	70.2
Sasaki et al., 2017	71.2
Radford et al., 2018	78.6

Implications

Proof of concept that IR could be useful

Current work:

- Machine learning components for query generation, antecedent selection
- Combine IR with deep learning
- Combine IR with reinforcement learning

Conclusions

Need for robustness in NLP systems

- Changing data distributions
- New events and entities
- Fine-grained reasoning about scenarios
- Implications for fairness and bias as well!

External knowledge as a way forward

- Challenge is to search for the right information from a vast collection
- IR + ML as a promising future direction

Acknowledgements

Thanks to my great collaborators:

- *Students:* Teng Long, Ali Emami, Emmanuel Bengio, Noelia de la Cruz, Ryan Lowe
- Doina Precup (McGill)
- Adam Trischler (MSR-Maluuba)
- Kaheer Suleman (MSR-Maluuba)

Funding sources:

- NSERC
- MSR