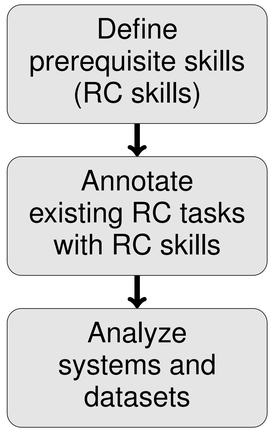
EMNLP 2016 Workshop: Uphill Battles in Language Processing, Poster #2 Austin, Texas, USA / November 5, 2016

An Analysis of Prerequisite Skills for Reading Comprehension

Saku Sugawara (University of Tokyo), Akiko Aizawa (National Institute of Informatics, Japan)
sakus@is.s.u-tokyo.ac.jp

Overview

- Challenge
  Evaluation method for reading comprehension (RC)
- Our goal
  Construct a general evaluation methodology that decomposes the RC process and elucidates:
  * the detailed performance of RC systems
  * the characteristics of RC tasks
  from multiple points of view: prerequisite skills.
- Our approach
  1. **Define a set of prerequisite skills** that are required for understanding documents
  2. **Annotate questions of an RC task** with the skills
  3. **Analyze the performances of RC systems** for the annotated questions to grasp the differences and limitations of their individual performances



Prerequisite Skills and Annotation

- 10 prerequisite basic skills were defined by investigating NLU tasks (WSC, COPA, CoNLL 15st, bAbI, and so on).
- We manually annotated questions with the RC skills that are required to answer each question (multi-labeling).
- We assume that when RC systems use RC skill, they already have the capability to recognize the facts described in the clauses that the skill pertained to.
- An example requires no skills:
  - Context: *Todd lived in a town.*
  - Question: *Where did Todd live?* — Answer: *In a town*

Reading Comprehension Skills and Annotation Results: accuracies and frequencies in RC tasks

RC skills (_* : "understanding of")	MCTest accuracy 1. Baseline SW+D	2. Smith No RTE	3. Smith RTE	MCTest Frequency	SQuAD Frequency	Description or Examples
List/Enumeration	51.1%	65.1%	61.9%	14.7%	5.0%	Tracking, retaining, and list/enumeration of entities/states
Mathematical operations	20.0%	30.0%	30.0%	1.6%	0.0%	Four basic operations and geometric comprehension
Coreference resolution	52.5%	63.6%	62.1%	63.8%	6.2%	Detection and resolution of coreferences
Logical reasoning	100.0%	75.0%	66.7%	0.9%	0.0%	Induction, deduction, conditional statement, and quantifier
Analogy	-	-	-	0.0%	0.0%	Trope in figures of speech, e.g., metaphor
Spatiotemporal relations*	48.9%	66.9%	67.1%	27.5%	2.5%	Spatial and/or temporal relations of events
Causal relations*	45.7%	62.0%	60.9%	14.4%	6.2%	Why, because, the reason, etc.
Commonsense reasoning	44.0%	61.3%	59.6%	41.9%	86.2%	Taxonomic/qualitative knowledge, action and event change
Complex sentences*	50.0%	65.9%	64.0%	20.6%	20.0%	Coordination or subordination of clauses
Special sentence structure*	46.2%	69.2%	73.1%	8.1%	25.0%	Scheme in fig of speech, constructions, and punct. marks
(Overall accuracy)	50.9%	66.2%	65.9%	-	-	Development sets: 120 (MC160) + 200 (MC500) questions

1: Richardson+ (2013)'s sliding window and word distance system (baseline), 2&3: Smith+ (2015)'s lexical matching systems (+RTE)
MCTest (2013): 320 questions (MC160+500 development sets), SQuAD (2016): 80 questions from the development set (v1.1)
Inter-annotator agreement: 85% for sampled 80 questions in MC500 development set

Annotation Example in MCTest (required 5 skills)

ID: MC160.dev.29 (1) multiple:
C1: The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping.
C2: She wandered out a good ways.
C3: Finally she went into the forest where there are no electric poles but where there are some caves.
Q: Where did the princess wander to after escaping?
A: Forest

Coreference resolution
- *She* in C2 = *the princess* in C1
- *She* in C3 = *the princess* in C1

Temporal relation
- the actions in C1 → *wandered out* in C2 → *went into ...* in C3

Commonsense reasoning
- *escaping* in Q ⇒ the actions in C1
- *wandered out* in C2 and *went into the forest* in C3 ⇒ *wander to the forest* in Q and A

Complex sentence and **special sentence structure** (ellipsis)
- C1 = *the princess climbed out ...* and [*the princess*] *climbed down ...*

Result: numbers of skills required in each question

# skill(s)	0	1	2	3	4	5	
MCTest Freq.	10.3%	28.4%	28.4%	23.8%	8.1%	0.9%	difficult?
SQuAD Freq.	5.0%	48.8%	37.5%	6.2%	2.5%	0.0%	easy?

Analyses and Conclusion

System analysis
A. All systems are still not good at *coreference resolution* and *commonsense reasoning*. (ideally the weakness is derived from the difference between accuracies of skill combinations...)
B. We could not observe that adding RTE significantly increased accuracy by small annotations :(

Dataset analysis
C. These scores reflect the difficulty of the datasets (SQuAD: Wikipedia (for adults); MCTest: tales (for children))
D. SQuAD has simple questions (mostly paraphrases).

⇒ Need more systems, datasets, and annotations!!
⇒ Are the skills sufficient? e.g. *commonsense reasoning...*