BUGS in Writing Genesese

Yang Zhutian (Yang)

February 8, 2018

Abstract

I care about why the Genesis story-understanding system cannot read longer and more varieties of stories. First, I research into the mechanism and challenges in parsing. Then, I test the Genesis English understanding ability by experimenting with the **Test translator/ generator** in `TheGenesisSystem.java`. Lastly, I discuss what can be done to make the Genesis system to understand more English written by people for people. In this report, I note down the concepts that I do not understand in red!70.

Contents

1 START **3**

1.1 Understanding English . . . 3

1.2 Answering Questions . . . 3

2 Parsering **4**

2.1 Latest Works . . . 4

2.2 Challenges . . . 4

2.2.1 "Garden Path" Sentences . . . 4

2.2.2 By Commonsense Knowledge . . . 4

2.2.3 By Mental Repairing . . . 4

2.2.4 TODO . . . 4

2.3 TODO . . . 4

3 Short Circuit **5**

3.1 What Short Circuit Can Do . . . 5

3.1.1 Grammatical . . . 5

3.1.2 Being Concise . . . 5

3.1.3 Being Clear . . . 5

3.2 What Short Circuit Cannot Do . . . 5

3.3 Where Short Circuit Makes Mistakes . . . 5

4 Via Genesis **6**

4.1 What Genesis Can Do . . . 6

4.1.1 Grammatical . . . 6

4.1.2 Structural . . . 6

4.1.3 Being Concise . . . 6

4.1.4 Being Clear . . . 7

4.2 What Genesis Cannot Do . . . 7

4.2.1 Wordings . . . 7

4.3 Where Genesis Makes Mistakes . . . 7

4.3.1 Preposition . . . 7

4.3.2 Others . . . 7

5 Interesting Examples **8**

5.1 Simple Sentences . . . 8

5.1.1 Missing Subject . . . 8

5.1.2 Confusing Article . . . 8

5.1.3 "there are" . . . 8

5.2 Complex Sentences . . . 8

5.2.1 Time Clauses . . . 8

5.3 Over Simaplify . . . 9

5.3.1 Specific Becomes General . . . 9

5.3.2 "it is ... to" . . . 9

5.3.3 "it is ... that" . . . 9

5.4 Over Interpret . . . 9

5.4.1 "should be" . . . 9

5.4.2 "is fun" . . . 9

5.5 Awkward English . . . 10

5.5.1 "in spite of the fact that" . . . 10

5.5.2 "the fact that" . . . 10

5.5.3 "like somebody" . . . 10

5.5.4 "be like somebody" . . . 10

1 START

Genesese is English that is readable by the Genesis system. I first look into how START system works.

START (SynTactic Analysis using Reversible Transformations) system can *understand* English sentences, *generate* summarized English sentences in its inner language, and *answer questions* in English based on the summarized English sentences. (1997)

1.1 Understanding English

The START system *understands* an English sentence by building its syntactic structure.

1. It breaks up the English sentence → *kernel sentences* which contains only one verb. example of a kernel sentence? what is its relationship with a parse tree?
2. It builds a representational structure containing all syntactic *parameters* of the sentence.
   - The salient parameters—**subject, object, and their relation**—are rearranged → *ternary expressions (T-expressions)*, having the form `<subject relation object>`.
   - Some parameters—**adjectives, possessive nouns, prepositional phrases**—are also rearranged → T-expressions with special `relation` words like `relate-to`
   - The remaining parameters—**adverbs, tense, auxiliaries, voice, negation**—are recorded in *history* which will not be included in the summarized English sentences. Are they not in the knowledge base? for example, if the sentence is "I bought flowers to Ben happily," then I can not ask if I was happy?

     auxiliaries: i.e., do, have, will, shall, would, should, can, could, may, might, must, ought
     voice: i.e., positive, negative ("to be done?")
3. The T-expressions together form the *knowledge base* which is like a "digested summary" of the original English sentence.

Input	Knowledge base
Bill surprised Hillary with his answer	`«Bill surprise Hillary> with answer>` `<answer related-to Bill>`

Table 1: An example of a sentence's knowledge base that contains two T-expressions.

Type	Input questions	Formulated questions
wh-question	Whom did Bill surprise with his answer?	`«Bill surprise whom> with answer>`
yes-no question	Did Bill surprise Hillary with his answer?	`«Bill surprise Hillary> with answer>`

Table 2: An example of a sentence's knowledge base.

1.2 Answering Questions

The START system *answers questions* about an English sentence by *formulating* the questions → the same structure as the T-expressions of the sentence and *matching* in the knowledge base.

2 Parsering

2.1 Latest Works

Parsey McParseface (2016) scans the words in order, guesses their roles in the sentence, and has practiced with thousands of sentences prepared by linguists.

displaCy (2017) visualizes part of speech by drawing arrows.

[TODO] Try Stanford

2.2 Challenges

2.2.1 "Garden Path" Sentences

The following sentences took you to a garden-path:

1. The old man the boat.
2. I convinced her children are noisy.
3. The coach smiled at the player tossed the frisbee.
4. The cotton clothes are made up of grows in Mississippi.
5. The horse raced past the barn fell.

While scanning each sentence from left to right, you might wander, for example, why "fell" appears at the end of Sentence # 5 and screw up the whole meaning. Actually, the right to look at it is:

> The horse (that was raced past the barn) fell.

2.2.2 By Commonsense Knowledge

Experiments have shown that we understand garden-path sentences using the context and our experience. We bias toward the most common word patterns and throw away unlikely interpretations. For example:

1. While the man hunted the dear ran into the woods.
2. While the man hunted the vice president ran into the woods.
3. My dad mixed the batter with the blueberries.

Sentence # 2 is easier to read than # 1 because we know the president is rarely hunted.

We understand that the blueberries are mixed together with the batter because we know that they are rarely used as a tool to mix the batter.

By the way, we humans also understand the grammar of a sentence faster if it has more familiar words.

2.2.3 By Mental Repairing

Roger Levy, a cognitive scientist now at MIT, thinks that our brain may try to repair sentences by predicting words that might have gotten lost, mumbled, or overlooked Levy (2008). This "fuzzy" recognition strategy helps us understand badly written sentences.

2.2.4 TODO

2.3 TODO

3 Short Circuit

To tell stories to the Genesis system, I should write Genesese that it will understand. Perhaps I could be a better communicator by writing Genesese that it will speak. Perhaps I could even cheat by being lazy and relying on the `Short circuit` to correct my English → what Genesese will speak.

3.1 What Short Circuit Can Do

3.1.1 Grammatical

- **Capital**: Short circuit capitalizes the first character.
- **Article**: Short circuit corrects your "a" and "an".
- **Subject-object agreement**: Short circuits corrects your third-person singular "-s".
- **Comma**: Short circuit adds comma after introductory phrases and clauses.

  e.g. in the morning

3.1.2 Being Concise

- **Space**: Short circuit deletes extra space
- **Adverb**: Short circuit deletes "please"
- **Needlessly Complex**: Short circuit rewrites "at the present (time)" → "now"

3.1.3 Being Clear

- **Negation**: Short circuit rewrites "n't" → "not"
- **Parallel structure**: Short circuit splits at "and," "or" and "but" using semicolons.
- **"it is that"**: Short circuit rewrites "it is ... that sb. do" → "it is ... for sb. to do"

3.2 What Short Circuit Cannot Do

Short circuit will respond "Please try again later" if there are more serious grammar mistakes

- **Tense**: "have came"
- **Missing "be"**: "while sad"
- **Missing "to"**: "cause me do"
- **"not only ... but also ..."**
- **"the reason that"**

3.3 Where Short Circuit Makes Mistakes

- **"early in the morning"**: Short circuit rewrites "early in the morning I ..." → "early, in the morning, I ..."

4 Via Genesis

4.1 What Genesis Can Do

4.1.1 Grammatical

- **Article**: Genesis adds "a" and "an" if you forget it.
- **Adverb**: Genesis puts adverbs to the end of sentence .

  e.g. rarely, often, never; here; greatly, clearly; already
- **Tense**: Genesis corrects the tense for you if there is violation of tense in your sentence.

  e.g. "claimed that he is"

4.1.2 Structural

- **"you"**: Genesis deletes "you" that works as the subject.
- **"cause"**: Genesis rewrites "AA causes BB" → "For BB nilled AA"
- **"in order to"**: Genesis puts the "in order to" to the beginning of the sentence.
- **Introductory phrases**: Genesis puts introductory phrases after the verb.

  e.g. "In the morning, I go to gym" → "I go in the morning to gym"
- **Conditional clause**: Genesis puts clauses to the end of sentence.

  eg. if, because. Problem with "since," and "when," and "while"

4.1.3 Being Concise

- **Adjective**: Genesis deletes certain adjectives.

  e.g. very, too, really, especially, extremely, well-earned; many of, some of
- **Redundancy**: Genesis deletes some of the redundant words.

  e.g. (increasingly) more, (other) choices
- **Zero phrases**: Genesis deletes phrases that say nothing.

  e.g. it is true, it is dangerous/ interesting to
  It is interesting to note that I love him.
  It is interesting to note that I love him.
  Notes my love for him.
- **"one's"**: Genesis deletes "one's"
- **"do you"**: Genesis deletes "do you," but not "do I" or "does he," in yes-no questions.

4.1.4 Being Clear

- **Parallel structure**: Genesis splits at semicolons to make multiple simple sentences.
- **"he ... he"**: Genesis rewrites the second "he" → "himself."
- **"he ... his"**: Genesis rewrites the "his" → "the."
- **"of"** as "belonging to": Genesis rewrites the "of" → possessive nouns.
- **"not at all"**: Genesis rewrites "not at all late" → "not late at all."
- **"that"** that marks the beginning of a declarative content clause: Sometimes Genesis deletes "that", sometimes it doe not. You will see by pressing "run" multiple times.

4.2 What Genesis Cannot Do

Genesis shows `Unable to generate text` if it cannot understand the wordings or the structure.

4.2.1 Wordings

- **"occur"**: It occurs to me that I cannot find my key. **"in terms of"**: The job is unattractive in terms of salary.

4.3 Where Genesis Makes Mistakes

Genesis makes mistakes when there is ambiguity in the input English.

4.3.1 Preposition

- **"for"**: Genesis rewrites "AA of BB" → possessive.
  - "for a period of time" → "for time's period."
  - "a great number of students" → "students' great number."(same with "none of")
  - "a great portion of students" → "Unable to generate text.A large portion."
  - "take possession of the reward" → "take the reward's possessions."
- **"to"**: Genesis treats "to" as purpose.
  - "it is a good way to use time" → "In order to use time, it is a good way."
  - "it is a pleasure to meet you" → "In order to meet you, it is a pleasure."
- **"in"**: Genesis treats the adverb before an adverb phrase as parallel.
  - "Early in the morning, I go to gym" → "I go in the morning to the gym early,"
  - "I go to gym early in the morning" → "I go to gym in the morning early."

4.3.2 Others

- **Tense**: "i know that it was raining" → "i know it is raining"
- **"being there"**: Genesis rewrites "while being there" → "while being a there"

5 Interesting Examples

5.1 Simple Sentences

5.1.1 Missing Subject

I am your friend.
I am your friend.
Am your friend.

She is your friend.
She is your friend.
Is your friend.

I am your love.
I am your love.
I am an ing love.

5.1.2 Confusing Article

She is a person's friend.
She is the person's friend.
She is person's friend.

5.1.3 "there are"

There were many leaves.
Many leaves were.
Leaves were.

5.2 Complex Sentences

5.2.1 Time Clauses

He forgot his key since being there.
He forgot his key since is there.
Unable to generate text.

He forgot his key since he was there.
He forgot his key since being there.
Unable to generate text.

He forgot his key when he was there.
He forgot his key when he was there.
Unable to generate text.

He forgot his key while being there.
He forgot his key while being there.
He forgot his key while he is a there.

5.3 Over Simaplify

5.3.1 Specific Becomes General

We should respect one's time.
We should respect time.
We respected time.

5.3.2 "it is ... to"

It is good to use time.
It is good to use time.
Uses time.

I told him that it is dangerous to go there.
I told him that it is dangerous to go there.
I told him that go there.

5.3.3 "it is ... that"

It is dangerous that I love you.
It is dangerous for me to love you.
I love you.

It has rarely been the case that I speak Japanese.
It has rarely been the case fore me to speak Japanese.
Speak has been a case.

5.4 Over Interpret

5.4.1 "should be"

It should be pointed out that I love him.
My loving him should be pointed out.
My loving for him was pointed out.

I do not know if I should be there.
I do not know if I should be there.
I do not know if I was there.

5.4.2 "is fun"

He did not think that studying Latin was fun.
He did not think that studying Latin was fun.
He did not think to study Latin/ He did not think (that) studied Latin.

5.5 Awkward English

5.5.1 "in spite of the fact that"

I went there in spite of the fact that it was raining.
I went there in spite of there's raining.
Unable to generate text.

5.5.2 "the fact that"

I know the fact that it was raining.
I know about **it**'s raining.
I know that it rains.

The fact that it is hard bothers me.
It bothers me that it is hard.
It bothers me for it to be hard.

5.5.3 "like somebody"

I do not know if I like him.
I do not know if I like him.
null.

I do not know if I hate him.
I do not know if I hate him.
I do not know if I hate him.

5.5.4 "be like somebody"

I do not know if I am like him.
I do not know if I am like him.
I do not know if I am. I am like him.

References

(1997). URL: http://people.csail.mit.edu/boris/webaccess/node1.html.
(2016). URL: https://github.com/tensorflow/models/tree/master/research/syntaxnet.
(2017). URL: https://demos.explosion.ai/displacy/.
Levy, Roger (2008). "A noisy-channel model of rational human sentence comprehension under uncertain input". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 234–243.