Why is it Hard to Understand Original English Questions?

Working Note 32

Peter Clark and John Thompson
Boeing Phantom Works

January 2009

peter.e.clark@boeing.com

Abstract

The AURA system is currently unable to reliably understand original English AP questions, and instead the user has to reformulate them into the controlled language CPL. The purpose of this report is to characterize why the original English is so difficult to process. Clearly, the need for world knowledge is part of the challenge. However, another significant challenge is crossing the language-knowledge gap, namely mapping from the many ways of saying something in English to the correct representation in the target ontology. This requires solutions to both linguistic and semantic processing issues, and (possibly interactive) algorithms for exploiting available world knowledge. We concentrate on this second challenge here, and attempt to characterize the different types of "mismatch" that create this gap between language and knowledge in the first place. This note thus complements an earlier report from Cycorp, and examines the same example questions discussed in that report.

1. Introduction

The purpose of this report is to characterize why original English AP questions are so difficult to process, and to provide some categories for the types of problems which arise. It complements an earlier report from Cycorp on the utility of common sense knowledge for question understanding (Curtis et al, 2008), and examines the same example questions in that earlier report.

One can consider the challenges of interpreting the original English questions in two broad classes:

1. **Linguistic Challenges:** Performing a correct syntactic analysis

   These challenges are concerned with parsing the original sentences correctly and creating a correct, initial "logical form" or semantic representation of the text. All the usual linguistic challenges fall into this category, in particular: prepositional phrase attachment, pronoun resolution, reference resolution, compound noun interpretation, word sense disambiguation, semantic role labeling, proper noun interpretation, quotations, parentheticals, units of measure, comparatives, negation, temporal reference, and ellipsis. Although they are here labeled as "linguistic," they can be guided by semantic knowledge from a KB and/or feedback from downstream processes, and some (e.g., word sense disambiguation) involve both linguistic and semantic issues.

2. **Semantic Challenges:** Crossing the Language-Knowledge Gap

   Broadly speaking, this involves mapping the many ways of expressing something in English into meaningful representations in the KB. For example, the phrase "straight up" should be mapped in AURA to "direction of 90 degrees." This challenge is formidable as:

- there are many ways of expressing the same idea in English
- there is often a mismatch between the way the problem is conceptualized in the original English and the way it can be understood by AURA
- the original English may be "sloppy" (i.e., not precisely correct), requiring some coercion to find the intended meaning.

This note concentrates mainly on the semantic challenges but also makes a few comments about the linguistic challenges.

2. Linguistic and Semantic Challenges

2.1 Linguistic Challenges: Performing a correct syntactic analysis

The initial language processing tasks (parsing and related activities) are quite challenging but not insurmountable. Our full language processor (BLUE - Boeing Language Understanding Engine, (Clark and Harrison, 2008)) often identifies the correct syntactic structure for the original English sentences that the Cyc report examined. While there are numerous linguistic challenges (as listed earlier), we mention just a few below that are interesting in the Cyc examples.

(a) Attachment Preferences

There is scope for using world knowledge to help decide where prepositional phrases should be attached. For example, in:

(1) A shell is fired into the air with an initial velocity of 300 m/s

a system should prefer the parse:

(2) A shell is fired (into the air) (with an initial velocity of 300 m/s)

rather than:

(3) A shell is fired (into (the air with an initial velocity of 300 m/s))

i.e., the velocity is of the firing event rather than the air. This example is particularly challenging to disambiguate as "velocity" is a valid property of both firing and air. Rather, pragmatic and contextual knowledge is needed to tell the system that velocity is more typically associated with firing events (particularly in the context of physics problems).

(b) Coreference

(4) A ball is thrown. What is the acceleration [of the ball, when it is] at a height of 2m during the movement [of the ball] upwards?

Note the ellipsis (missing arguments) above. To resolve references to "the acceleration" and "the movement," one needs to infer the presence of a Move event after the Throw. The Cyc report shows Cyc has an axiom for this inference (ProjectilePropeling —causes→ ProjectileMotion). CLib also has a similar axiom (Throw —subevent→ Propel). The challenge here is not only having the world knowledge, but also the language interpretation mechanism to introduce it appropriately in the scene to resolve this reference. This problem is sometimes referred to as "indirect anaphora resolution." (Fan et al, 2005).

A second example is:

> (5) Gravity stops the shell briefly before it falls back down.

The Cyc report nicely discusses the task of identifying "it" to be "the shell" (rather than "gravity"), using world knowledge that shells (but not gravity) can fall. CLib also has this knowledge (moving objects can only be Tangible-Entities).

(c) Ellipsis

There are several cases of ellipsis (missing arguments) in the questions. Some are shown below, with the missing arguments shown in square brackets:

> (6a) thrown from a height of 1 m [above ground]
>
> (6b) Before cell division in mitosis but afterward [after cell division] in meiosis
>
> (6c) Which choice shows the correct pairing? (B) Adenine [pairs] with cytosine.

Identifying the missing argument requires a mixture of linguistic knowledge (e.g., "before X and after" means "before X and after X"), and pragmatic knowledge (e.g., heights are by default above the ground.)

2.2 Semantic Challenges: Crossing the Language-Knowledge Gap

The biggest problem with the original English is crossing the language-knowledge gap. The problem occurs when what is written in the original English does not directly fit into the target ontology. The reasons the KB may not be able to accept the original conceptualization are because the original English may not have a complete mapping to the target ontology (e.g., a word does not map to any concept), or a literal interpretation violates constraints in the ontology, or the knowledge is inexpressible in the KB's ontology or in the underlying KR language

As a first-cut characterization, there are three types of semantic "mismatch" between language and knowledge that can occur:

I. **Paraphrase:** there are many ways of expressing the same idea in English

II. **Conceptual:** The original way of conceptualizing the problem is inexpressible, incomprehensible, or unusable in AURA, and so needs to be recast in a different way.

III. **Non-Literal:** A literal interpretation of the English is incorrect, and so needs to be coerced to match the KB.

We illustrate these below.

I. Paraphrase mismatch

An example of the "paraphrase" problem is the different ways of saying "up," as discussed in the Cyc report:

> (7a) "thrown directly upwards" → direction 90 degrees
>
> (7b) "fired straight up" → direction 90 degrees

The challenge here is not in creating any specific mapping (this is straightforward), but accommodating the thousands of possible variations that can occur. Within a limited domain it might be possible to author these by hand, but for broader coverage, machine learning techniques are likely required (there is a subfield of AI devoted specifically to paraphrase learning).

II. Conceptual mismatch

As a first-cut, we can identify four types of conceptual mismatch:

IIa. Properties attached differently (approximately, metonymy)
IIb. Different event conceptualization
IIc. Different object conceptualization
IId. Different relation conceptualization

We illustrate these here:

IIa. Properties attached differently

As an example, consider a thrown object. If an object is thrown, we might refer to velocity as a property of:

- the object ("the velocity of the object is...")
- the throwing event ("thrown with a velocity of...")
- the movement event following the throw ("it travels at a velocity of ...")

However, in AURA velocities are properties of movement events. As a result, the language interpreter needs to move the velocity property from its placement in the original English to the appropriate movement event in the final representation.

These transformations can become quite complex, in particular when they interact with other linguistic challenges (e.g., indirect anaphora, metonymy elsewhere). CPL has a small number of hand-coded transformation rules for handling such cases, but this is a rather piecemeal solution to the problem.

IIb. Different event conceptualization

> (8) After traveling up, gravity eventually stops the shell briefly (velocity is zero) before it falls back down. How high does the bullet go?

The above text suggests three events (travel up, stop, fall down). However, AURA expects one event (a projectile movement). Somehow the three need to be coerced into one. It also requires world knowledge that the maximum height is the height when the shell stops (= the end of the travel up event), and so one could recast this problem by simply looking at the "travel up" event and ask for its distance. In fact, this is exactly what the SME did in the example CPL in the Cyc report:

> There is a move.
> The initial y speed of the move is 330 m/s.
> The final y speed of the move is 0 m/s.
> What is the distance of the move?

Automating this, though, would clearly be very hard. This is a good example of the challenges of the language-knowledge gap.

A second example is:

> (9) What is the acceleration [of the ball] at a height of 2m during the [ball's] movement upwards?

The English is asking for the property (acceleration) of the ball at a position during movement. This is not representable in AURA because CLib doesn't support asking for acceleration at a position or time-point; rather, one can only ask for acceleration at the start (initial-acceleration), throughout (acceleration), or end (final-acceleration) of a movement. So to coerce this question into CLib, it needs to be conceptualized as a movement which *starts* with the throw at 1m and *ends* with the ball at 2m, i.e., it is a 1 m vertical movement. This is quite a change from the original!

IIc. Different object conceptualization

A similar phenomenon can exist for objects (although there are no obvious examples in the Cyc report). For example, a biology question might state "the cell is surrounded by a protective membrane," while the KB might consider the protective membrane to be *part of* the cell. Such incompatibilities need to be resolved to avoid (here) two membranes appearing in the interpretation.

IId. Different relation conceptualization

The Cyc report highlights a nice example of this, where the biology-specific complement(x,y) relation (between complementary base pairs) is instead referred to as "pairing" in the English, thus appealing to an event (X pairs with Y) rather than the state (complement) resulting from that event:

> (10) Which choice shows the correct pairing of the nitrogen bases in DNA?

To understand the English, one would need to relate "pairing" to complement(x,y). While one might simply add "pairing"(n) as a synonym for "complement," a more general solution would be to relate the verb "pair"(v) also, so that verbal derivatives can be handled, e.g., "Adenine pairs with cytosine,, "Adenine is paired with cytosine," "The pairing of adenine and cytosine," "Adenine and cytosines are base pairs."

In many cases, adding this mapping is straightforward; again, the challenge is that large numbers of such mappings may be needed.

III. Non-Literal

Sometimes what is written does not make sense when taken literally. Again, as discussed in the Cyc report:

> (11) A small artillery shell is fired....How high does the bullet go?

Although "the bullet" was intended to refer to "the shell," a bullet is strictly not a shell, and so some algorithm is needed to realize that coreference is intended, and to locate the coreferential object.

A second example is:

> (12) After traveling up, gravity eventually stops the shell briefly.

"briefly" suggests a (small) time interval; however, in a strict sense there is no interval, as the "stop" denotes a time *point*. Thus the system would have to somehow avoid associating a non-zero time interval to the stop event.

3. Conclusion

This report has presented a first-cut characterization of challenges in understanding original English AP questions. The linguistic challenges are difficult, but in many cases our language engine can produce a good syntactic analysis of the original already, and so they may be surmountable. The semantic challenges pose a larger difficulty, when the original conceptualization does not fit the target ontology. In some cases a simple transformation (e.g., a paraphrase or substitution) is sufficient to cross the gap, but in other cases a more radical re-conceptualization of the problem is needed. In either case, though, there is a critical need for algorithms to make such transformations, possibly using interaction and dialog with the user, to making progress towards handling the original English and exploiting world knowledge that might be available from other sources.

References

Clark, P., Harrison, P. 2008. Boeing's NLP System and the Challenges of Semantic Representation. In Proc *SIGSEM Symposium on Text Processing (STEP'08),* Venice, Italy.

Curtis, J., Lenat, D., Shepard, B., Witbrock, M. 2008. *Preliminary Analysis of the Utility of Common Sense, Domain, and Lexical Knowledge to the Formalization of Advanced Placement Exam Questions*. Technical Report, Cycorp.

Fan, J., Porter, B., Barker, K. 2005. Indirect Anaphora Resolution as Semantic Path Search. In: *Proc KCap'05.*