Precise Understanding of Language by Computers

Iddo Lev

October 20, 2005

project website:
http://www.stanford.edu/~iddolev/pulc

[This document is still a draft]

1 The Need

Humans are able to learn a natural language like English when they are very young thanks to two things: the innate structure of their brains, and the fact that they are situated agents that act in the real world and interact with their environment, thereby learning the meaning of words and sentences and their proper use in social contexts. Computers, in contrast, lack both of these things, and until they have them, the only way they can come to understand language (to some degree) is if they have a lot of knowledge about language and about the world.

The knowledge about the world can be *anything*, ranging from simple facts, to commonsense knowledge (such as if you drop a cup it will fall to the floor and break), to expert knowledge in particular domains. In contrast, the knowledge about language, while vast, is relatively much smaller.

I think that we need a project on the scale of the Human Genome Project. Just as this latter project aimed to map out the entire human DNA sequence, the Precise Understanding of Language by Computers (PULC) project would aim to map out the entire human knowledge of the English language.[1] Of course, the entire field of Linguistics aims at mapping out the human knowledge of English and all other natural languages, but the emphasis of PULC is to state the knowledge in a comprehensive, consistent, and coherent format, so precise that it could be used by a computer. There would be a PULC for other languages as well, but here we'll only talk about English.

2 Previous Work

In the past several decades there has been a decent progress in discovering and documenting this knowledge in Linguistics. But there are still many gaps in this

[1] Just as the Human Genome Project aims at a "generic" human DNA sequence that is shared by most humans, so PULC aims at a conventional knowledge of English that is shared by most English speakers.

knowledge. Moreover, the knowledge is usually stated only semi-rigorously, and is not immediately applicable for a computational system.

There is of course also a lot of work in computational linguistics, or more precisely, natural language processing (NLP), that aims at producing computer applications that do useful things with language input. But just as linguists to the most part ignore computational issues, NLP researchers by and large ignore linguistic theory (people in NLP both aim at immediate practical applications and are usually engineers that lack – and are even uninterested in – linguistic knowledge).

3 We Need a Clear Goal and Plan

My lack of knowledge of genetics might cause me to simplify the picture, but as far as I understand, the Human Genome Project had a clear goal and a clear methodology from the outset. Certainly, not all details were known, but nonetheless, what the outcome should look like was pretty much clear – a sequence of A,C,G,T letters that corresponds (in precise ways) to sequences of DNAs extracted from humans. And although technological innovation throughout the project allowed for gradually increased speed, efficiency, and accuracy of the mapping, the basic method of how to map the genome was clear throughout, and people knew they merely needed to sit and apply these methods section by section to the DNA.

The situation with PULC is more complicated for various reasons. It is not clear what the goal is, nor how to get it.

The first distinction to be made is that the focus of the project is only Linguistic Knowledge. That means we need to draw a borderline, as clear as we can, between that knowledge and the enormous extent of general commonsense knowledge. The border is sometimes fuzzy, e.g. it is not immediately clear whether the knowledge that X bought Y from Z implies that Z sold Y to X is world knowledge or linguistic knowledge, and similarly the fact that if John, Mary, and Sue sneezed then each of them did so individually, but if they gathered in the street, they did so as a group. But we need to make such decisions as much as possible.

4 The Goal

4.1 Linguistic Knowledge

What is of interest to us is hard to define precisely, but what I have in mind is knowledge of roughly the following kind:

- Morphology: this is pretty much well known and a solved issue.
- Structural Knowledge:
  - Syntax, syntax-semantics interface, structural semantics: cataloging the different structural forms in English sentences (e.g.: basic grammar categories and rules, wh-extraction, various linguistic operators, extraposition, anaphora, ellipsis, respectively-constructions, etc.), and their corresponding meaning representations (taking into account e.g. ambiguities

regarding scope-taking operators and plurality). This knowledge should be comprehensive enough to completely specify all the possible meaning representations that we would possibly want to get for an English sentence, out of context.

  - Non-truth-conditional semantics: presupposition triggers and rules (e.g. from definite expressions and attitude verbs), and triggers for sentence-based implicatures (from logical operators, and scalar implicatures from quantifiers).
  - Discourse knowledge: how truth-conditions from adjacent sentences can be combined, interaction with anaphora and presuppositions, discourse relations and their triggers (such as: explanation, elaboration, justification)

- Lexical Knowledge:

  - Collocations: lexical phrases with limited compositionality, such as idioms, stock phrases, preferences for N-N and Adj-N combinations
  - Lexical and lexical-semantic foundations of syntax: connection between grammatical functions, subcategorization frames, thematic and semantic roles
  - Other? (tense and aspect? connection between 'buy' and 'sell' – is this linguistic knowledge or world knowledge?)

  The border needs to be defined more clearly for the lexical knowledge. E.g. is the connection between 'kill' and 'die' part of linguistic or world knowledge? The answer would depend on the scope of such knowledge. If it is quite restricted and has an important effect on linguistic form, it should be considered part of linguistic knowledge. If it is very wide and unbounded, and has little effect on linguistic form, it should be considered part of world knowledge.

As far as additional pragmatic and other knowledge, such as invited inference, non-grammar-based implicatures, Gricean maxims of conversation, etc., it is currently hard to assess how much of this knowledge is linguistic and how much is just general commonsense inference. Currently we'd say this is application-based.

4.2 Computational Apparatus

The machinery that takes the input text and, using the linguistic knowledge, translates the text to a meaning representation.

4.3 Results and Evaluation

When all this knowledge is available in the computer, what we'll be able to get is a program that translates English texts to logical representations of the literal meaning(s) of the text. If this information is then combined with a logical representation of relevant general world knowledge, inference could be made from the information conveyed by the text. In those areas where the required extra world knowledge is relatively small or easily formalizable (e.g. when solving logic puzzles, and even math

problems), the text could actually be completely automatically understood (for the purpose of the task they intend to serve).

We need criterions to know to what extent our goal is achieved. The criterions can be expressed as a series of tests (along the line of (FraCaS, 1996)) that check whether the computer understands the meaning of certain constructions, i.e. can correctly make logical inferences based on the information they convey. There will be a (long) list of linguistic phenomena that the computer should know about, and tests for each.

In contrast to currently prevailing evaluation measures in NLP, the tests should be per construction (or per a set of constructions) rather than per-corpus, because what we want to check is whether the computer has correct and complete linguistic information about English, not whether it has a "broad coverage" but shallow level of understanding of language that allows it to produce some sort of low quality representation for an arbitrary text with arbitrary linguistic phenomena. We aim at high-quality, precise, and deep semantic understanding.

5 The Methodology

We want to have a methodology that, just as in the Human Genome Project, can be systematically applied to one linguistic phenomenon after another, to map the linguistic knowledge of English into a computational form.

In the Human Genome Project, the goal is the map the entire DNA sequence completely, each and every bit of it. There is no expectation that mapping just part of it is going to allow us, by any means whatsoever, to infer the rest. DNA is like a computer program, it has many different parts, and although there are repetitions and redundancies, there is no guarantee that knowing even 99% of it could allow us to infer anything *of real value* about the other 1%.

The same is true, I believe, about English. For example, knowing everything about basic grammar, wh-extraction, and anaphora does not tell one anything at all about ellipsis. Therefore, what is needed is a complete and systematic study of the entire range of phenomena of the language.

To sharpen this point: if a text employs a particular linguistic construction (e.g. reciprocals), and this construction is used to convey an important part of the information in the text, then the computer must have knowledge about this construction, or else it will not be able to produce a correct logical representation of the text's meaning.

The proposed methodology is to systematically map one linguistic phenomenon after another. This means: finding the relevant knowledge in books and papers in Linguistics, formalizing it, extending it, and integrating it into a consistent and coherent whole. This effort is similar to the efforts in (Pollard and Sag, 1994; Copestake et al., 2002; XLE, 2002; Carpenter, 1998), except that:

- The effort would take into account not only syntax but also structural and lexical semantics.
- The knowledge would not only be implemented on a computer but would be published in its entirety in a non-computerized form, with comprehensive ex-

planations and documentations, so that human scholars can understand and modify it. This is just as important as the difference between computer code and its documentation.

- The fruits of the effort would be freely available and would include contributions from many people. This contrasts with previous proprietary efforts (such as (Alshawi, 1992) and at HP-labs in the '80s). The scope of the effort is much larger than any one company or research center can cope with alone (and the work produced by such bodies often remains unavailable to the public even years after it was discontinued).

This is an enormous effort, which will take a lot of time and work, but it is extremely valuable, and I think there is no other way to do it. Machine Learning technology cannot by itself solve the problem. Supervised learning requires an enormous effort of annotating corpora anyway (e.g. the Penn Treebank). Unsupervised learning is unlikely to produce any valuable knowledge as long as we humans have not yet figured out what semantic representations we would like to get for a text in order for the computer to achieve high-quality precise semantic understanding.

5.1 What's Involved

The process of uncovering structural linguistic knowledge goes roughly like this:

1. Pick a linguistic phenomenon and collect a few sentences exhibiting it. The sentences should be collected from natural sources and should represent as many different kinds of uses of the phenomenon as possible. It is important to keep the surrounding textual context for each sentence so it will be possible to know what the meaning was (out of the potentially many possible meanings of the sentence out of context).

2. For each sentence, write a logical formula that expresses the truth conditions of the sentence. The formula should be written in a formalization that has precise model-theoretic semantics and that can capture the way we conceptualize the meaning (see more on that issue in my draft "Inference from Natural Language Texts: Conceptualization, Representation, and Computability"). More than one formula may be written for a sentence if it can have different readings in different contexts.

3. For each piece of the logical formula, identify its source in the English sentence. Add rules to the already developed grammar which specify how the logical formula could be assembled via the syntactic grammar from the sentence. This can be done along the lines of (Lev, 2005). The addition should be consistent with the existing grammar, or require to make changes to it so that the result will be consistent.[2]

[2] "Consistent" means: if a linguistic phenomenon is associated with a certain formal construction, it should always be associated with it, rather than changing the construction whenever it doesn't fit with the rest of the grammar. To make this point more clear, look at a negative example in (Lev, 2005): the way the representation of a transitive verb changes in a Hole-Semantics-like framework depending on whether the verb has a normal direct object or a reciprocal.

6 Possible Objections and Discussion

[rather than say: what are the tools that are available to us today and let's see what we could do with them, if necessary combining them with some tricks and ad-hoc ideas; let's ask how would a system look like which is complete and precise and has all the knowledge it needs. And then see how to either obtain this information or approximate it based on a principled model only in those cases where approximation wouldn't hurt too much.]

References

Alshawi, Hiyan, ed. 1992. *The core language engine*. MIT Press.

Carpenter, Bob. 1998. *Type-logical semantics*. MIT Press.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan Sag. 2002. English Resource Grammar website. http://lingo.stanford.edu/erg.html.

FraCaS. 1996. Using the framework: Deliverable 16 of the FraCaS project. http://www.cogsci.ed.ac.uk/∼fracas/.

Lev, Iddo. 2005. The syntax-semantics interface: A gentle introduction to Glue Semantics (with some new results). Unpublished manuscript, Stanford University. http://www.stanford.edu/∼iddolev/.

Pollard, Carl, and Ivan Sag. 1994. *Head-driven phrase structure grammar*. Studies in Contemporary Linguistics. The University of Chicago Press.

XLE. 2002. Xle website. http://www2.parc.com/istl/groups/nltt/xle/.

project website:
http://www.stanford.edu/∼iddolev/pulc