An Alternative Roadmap to Deep Linguistic Intelligence

Merging Theoretical Linguistics and Deep Learning

By Sasson Margaliot
Jerusalem, Israel

Scientific Syntax

In the field of Deep Learning, work has begun in the direction of combining it with linguistics. For example, in the 2013 article by Socher et al., the process of building vectors is guided by the parse tree of a sentence, but everyone understands that much more needs to be done.

As of today, **real** theoretical linguistics has been completely ignored by the Deep Learning community. Consider the article by D'Alessandro, R. "The Achievements of Generative Syntax: a Time Chart", which is presently trending #1 on LingBuzz (a list of the most downloaded linguistic papers). This article lists 66 features that were reliably established by a group of researchers. The conclusions found therein were not arrived at easily. Over the past 60 years, thousands of professional linguists working in the framework of **theoretical linguistics**, have generated innumerable interesting ideas, and yet only a few made it onto this list. These features successfully withstood many years of critical evaluation, and are still believed to be valid.

This work of 60 years, partially summarized in the quoted article, is essentially the scientific description of Natural Language. These features should be included in the machinery of Natural Language Processing together with Deep Learning algorithms. Just as we do not attempt to "learn" exponential and hyperbolic functions using backpropagation algorithms, we shouldn't think that these fundamental features of Natural Language must be rediscovered through Deep Learning.

In contrast, whatever it is that is described by long-since-proven-wrong, archaic models of language used in Dependency Parsing and the like, is simply not Natural Language. Given the use of inadequate and outdated models of language, the recent efforts to solve the Natural Language problem by the Deep Learning community cannot possibly be considered sincere. It is as if the image-recognition researchers would voluntarily restrict themselves to only working with black and white, low resolution, unfocused pictures, while color, high resolution, focused pictures are readily available.

Need for Communication

There is no easy way to obtain access to the correct, most up-to-date syntactic theories. The researchers are presenting their work in a form that they alone can understand, peppered with heavy professional jargon, and make no effort for their work to be understandable by non-linguists, including computer scientists. In addition, various alternative proposals are always considered, and even a reader familiar with all the professional terminology still will not be able to easily figure out the current state-of-the-art in theoretical linguistics since theoretical

assumptions are shifting all the time. The goal of linguistic theorizing is to discover the truth regarding the workings of human language, not to make the work of computer scientists easier.

To build real, scientifically sound models of language, the artificial intelligence community will need to establish serious and productive dialog with the community of theoretical linguists. Computer scientists need to overcome their pre-conceived notions which they seem to have developed by reading the ideologically- motivated critics of "chomskian" linguistics, and start to receive guidance from those who know – the theoretical linguistic experts.

"Elephant Database"

Syntax is not the only problem. Attempts to apply "deep learning" to Natural Language are facing the inherent difficulty following from the fact that there is an intractable amount of information that forms the "knowledge of the world" (as opposed to the "knowledge of language") which nevertheless has to be available during language processing.

This information should not be "learned" by neural networks and placed in vectors – there is simply way too much of it. On the other hand, you cannot possibly expect the robots to be able to read all the various forms of databases. The only feasible solution is to have as much information as possible in one single and simple format. It turns out that the lion's share of world knowledge information (probably way over 99%) can be placed into one very simple database. I call it an "Elephant Database" because it is so simple that it doesn't require human specific intellectual capacity to work with it. The idea is to let robots work with astronomical amounts of information easily accessible in the "Elephant Database". Simply working with this information will not qualify this robot as a linguistic agent, but rather would be considered a pre-linguistic agent.

Physically this massive amount of information could be distributed all over the Internet, but in principle it can be visualized as just one huge Excel table with only four fields: (Class, ID, Class, ID). All of the records in this table always express implication. There is no need for documentation or any additional conventions. The definition of fields could be deduced from the actual content in the database.

This is pre-linguistic intelligence: there are no negations, disjunctions, and quantifiers. The ability to work with pre-linguistic intelligence is a necessary pre-condition for robots ultimately achieving linguistic intelligence.

Intelligent Agents

Human language doesn't exist outside of the context of a multi-agent environment. This is yet another necessary pre-condition for true language processing. The meaning of statements in Natural Language includes, importantly, the instructions for maintaining a database of presuppositions. This database (which is different from the pre-linguistic database described in

the previous section) contains information about what's going on in the present conversation, as well as information about other participants in the conversation.

Even the simplest word in English, 'a', needs the presupposition database in order to express its meaning. The function of the undefined article 'a' is to send the message that a new entry should be introduced in the presupposition database of the listening agent. In other words, to simply choose between 'a' and 'the', a robot needs information about the other robot's database of working presuppositions. The exchange of messages between linguistic agent that takes place during a conversation leads to the upgrading of each other's dynamic databases in real time.

It is most unlikely that the machinery of preposition databases would be "discovered" in the process of Deep Learning. Therefore it also has to be a part of the language processing engine, together with the scientific syntax algorithms, the pre-linguistic database layer , as well as Deep Learning algorithms.

References

Chomsky, N. 2013. "Problems of projections." Lingua 130. 33-49.

D'Alessandro, R. 2017 "The achievements of Generative Syntax: a time chart." lingbuzz

Fox, D. 2012 "Presupposition projection from quantificational sentences: trivalence, local accommodation, and presupposition strengthening."

Mikolov, T, A. Joulin and M. Baroni. 2016 "A Roadmap towards Artificial Intelligence"

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. 2013 "Recursive deep models for semantic compositionality over a sentiment treebank."