Last Words

Computational Linguistics and Deep Learning

Christopher D. Manning*
Stanford University

1. The Deep Learning Tsunami

Deep Learning waves have lapped at the shores of computational linguistics for several years now, but 2015 seems like the year when the full force of the tsunami hit the major Natural Language Processing (NLP) conferences. However, some pundits are predicting that the final damage will be even worse. Accompanying ICML 2015 in Lille, France, there was another, almost as big, event: the 2015 Deep Learning Workshop. The workshop ended with a panel discussion, and at it, Neil Lawrence said, "NLP is kind of like a rabbit in the headlights of the Deep Learning machine, waiting to be flattened." Now that is a remark that the computational linguistics community has to take seriously! Is it the end of the road for us? Where are these predictions of steamrollering coming from?

At the June 2015 opening of the Facebook AI Research Lab in Paris, its director Yann LeCun said: "The next big step for Deep Learning is natural language understanding, which aims to give machines the power to understand not just individual words but entire sentences and paragraphs."[1] In a November 2014 Reddit AMA (Ask Me Anything), Geoff Hinton said, "I think that the most exciting areas over the next five years will be really understanding text and videos. I will be disappointed if in five years' time we do not have something that can watch a YouTube video and tell a story about what happened. In a few years time we will put [Deep Learning] on a chip that fits into someone's ear and have an English-decoding chip that's just like a real Babel fish."[2] And Yoshua Bengio, the third giant of modern Deep Learning, has also increasingly oriented his group's research toward language, including recent exciting new developments in neural machine translation systems. It's not just Deep Learning researchers. When leading machine learning researcher Michael Jordan was asked at a September 2014 AMA, "If you got a billion dollars to spend on a huge research project that you get to lead, what would you like to do?", he answered: "I'd use the billion dollars to build a NASA-size program focusing on natural language processing, in all of its glory (semantics, pragmatics, etc.)." He went on: "Intellectually I think that NLP is fascinating, allowing us to focus on highly structured inference problems, on issues that go to the core of 'what is thought' but remain eminently practical, and on a technology

* Departments of Computer Science and Linguistics, Stanford University, Stanford CA 94305-9020, U.S.A. E-mail: manning@cs.stanford.edu.

1 http://www.wired.com/2014/12/fb/.

2 https://www.reddit.com/r/MachineLearning/comments/2lmo0l/ama_geoffrey_hinton.

doi:10.1162/COLI_a_00239

that surely would make the world a better place." Well, that sounds very nice! So, should computational linguistics researchers be afraid? I'd argue, no. To return to the *Hitchhiker's Guide to the Galaxy* theme that Geoff Hinton introduced, we need to turn the book over and look at the back cover, which says in large, friendly letters: "Don't panic."

2. The Success of Deep Learning

There is no doubt that Deep Learning has ushered in amazing technological advances in the last few years. I won't give an extensive rundown of successes, but here is one example. A recent Google blog post told about Neon, the new transcription system for Google Voice.[3] After admitting that in the past Google Voice voicemail transcriptions often weren't fully intelligible, the post explained the development of Neon, an improved voicemail system that delivers more accurate transcriptions, like this: "Using a (deep breath) *long short-term memory deep recurrent neural network* (whew!), we cut our transcription errors by 49%." Do we not all dream of developing a new approach to a problem which halves the error rate of the previously state-of-the-art system?

3. Why Computational Linguists Need Not Worry

Michael Jordan, in his AMA, gave two reasons why he wasn't convinced that Deep Learning would solve NLP: "Although current deep learning research tends to claim to encompass NLP, I'm (1) much less convinced about the strength of the results, compared to the results in, say, vision; (2) much less convinced in the case of NLP than, say, vision, the way to go is to couple huge amounts of data with black-box learning architectures."[4] Jordan is certainly right about his first point: So far, problems in higher-level language processing have not seen the dramatic error rate reductions from deep learning that have been seen in speech recognition and in object recognition in vision. Although there have been gains from deep learning approaches, they have been more modest than sudden 25% or 50% error reductions. It could easily turn out that this remains the case. The really dramatic gains may only have been possible on true signal processing tasks. On the other hand, I'm much less convinced by his second argument. However, I do have my own two reasons why NLP need not worry about deep learning: (1) It just has to be wonderful for our field for the smartest and most influential people in machine learning to be saying that NLP is the problem area to focus on; and (2) Our field is the domain science of language technology; it's not about the best method of machine learning—the central issue remains the domain problems. The domain problems will not go away. Joseph Reisinger wrote on his blog: "I get pitched regularly by startups doing 'generic machine learning' which is, in all honesty, a pretty ridiculous idea. Machine learning is not undifferentiated heavy lifting, it's not commoditizable like EC2, and closer to design than coding."[5] From this perspective, it is people in linguistics, people in NLP, who are the designers. Recently at ACL conferences, there has been an over-focus on numbers, on beating the state of the art. Call it playing the Kaggle game. More of the field's effort should go into problems, approaches, and architectures. Recently, one thing that I've been devoting a lot of time to—together with many other

3 http://googleblog.blogspot.com/2015/07/neon-prescription-or-rather-new.html.
4 http://www.reddit.com/r/MachineLearning/comments/2fxi6v/ama_michael_i_jordan.
5 http://thedatamines.com/post/13177389506/why-generic-machine-learning-fails.

collaborators—is the development of Universal Dependencies.[6] The goal is to develop a common syntactic dependency representation and POS and feature label sets that can be used with reasonable linguistic fidelity and human usability across all human languages. That's just one example; there are many other design efforts underway in our field. One other current example is the idea of Abstract Meaning Representation.[7]

4. Deep Learning of Language

Where has Deep Learning helped NLP? The gains so far have not so much been from true Deep Learning (use of a hierarchy of more abstract representations to promote generalization) as from the use of distributed word representations—through the use of real-valued vector representations of words and concepts. Having a dense, multi-dimensional representation of similarity between all words is incredibly useful in NLP, but not only in NLP. Indeed, the importance of distributed representations evokes the "Parallel Distributed Processing" mantra of the earlier surge of neural network methods, which had a much more cognitive-science directed focus (Rumelhart and McClelland 1986). It can better explain human-like generalization, but also, from an engineering perspective, the use of small dimensionality and dense vectors for words allows us to model large contexts, leading to greatly improved language models. Especially seen from this new perspective, the exponentially greater sparsity that comes from increasing the order of traditional word n-gram models seems conceptually bankrupt.

I do believe that the idea of deep models will also prove useful. The sharing that occurs within deep representations can theoretically give an exponential representational advantage, and, in practice, offers improved learning systems. The general approach to building Deep Learning systems is compelling and powerful: The researcher defines a model architecture and a top-level loss function and then both the parameters and the representations of the model self-organize so as to minimize this loss, in an end-to-end learning framework. We are starting to see the power of such deep systems in recent work in neural machine translation (Sutskever, Vinyals, and Le 2014; Luong et al. 2015).

Finally, I have been an advocate for focusing more on compositionality in models, for language in particular, and for artificial intelligence in general. Intelligence requires being able to understand bigger things from knowing about smaller parts. In particular for language, understanding novel and complex sentences crucially depends on being able to construct their meaning compositionally from smaller parts—words and multi-word expressions—of which they are constituted. Recently, there have been many, many papers showing how systems can be improved by using distributed word representations from "deep learning" approaches, such as word2vec (Mikolov et al. 2013) or GloVe (Pennington, Socher, and Manning 2014). However, this is not actually building Deep Learning models, and I hope in the future that more people focus on the strongly linguistic question of whether we can build meaning composition functions in Deep Learning systems.

5. Scientific Questions That Connect Computational Linguistics and Deep Learning

I encourage people to not get into the rut of doing no more than using word vectors to make performance go up a couple of percent. Even more strongly, I would like to

6 http://universaldependencies.github.io/docs/.
7 http://amr.isi.edu.

suggest that we might return instead to some of the interesting linguistic and cognitive issues that motivated noncategorical representations and neural network approaches.

One example of noncategorical phenomena in language is the POS of words in the gerund *V-ing* form, such as *driving*. This form is classically described as ambiguous between a verbal form and a nominal gerund. In fact, however, the situation is more complex, as *V-ing* forms can appear in any of the four core categories of Chomsky (1970):

		V +		V −	
N	+	Adjective:	an *unassuming* man	Noun:	the *opening* of the store
	−	Verb:	she is *eating* dinner	Preposition:	*concerning* your point

What is even more interesting is that there is evidence that there is not just an ambiguity but mixed noun–verb status. For example, a classic linguistic text for being a noun is appearing with a determiner, while a classic linguistic test for being a verb is taking a direct object. However, it is well known that the gerund nominalization can do both of these things at once:

(1) The not observing this rule is that which the world has blamed in our satorist. (Dryden, *Essay Dramatick Poesy,* 1684, page 310)

(2) The only mental provision she was making for the evening of life, was the collecting and transcribing all the riddles of every sort that she could meet with. (Jane Austen, *Emma,* 1816)

(3) The difficulty is in the getting the gold into Erewhon. (Sam Butler, *Erewhon Revisited,* 1902)

This is oftentimes analyzed by some sort of category-change operation within the levels of a phrase-structure tree, but there is good evidence that this is in fact a case of noncategorical behavior in language.

Indeed, this construction was used early on as an example of a "squish" by Ross (1972). Diachronically, the *V-ing* form shows a history of increasing verbalization, but in many periods it shows a notably non-discrete status. For example, we find clearly graded judgments in this domain:

(4) Tom's winning the election was a big upset.

(5) ?This teasing John all the time has got to stop.

(6) ?There is no marking exams on Fridays.

(7) *The cessation hostilities was unexpected.

Various combinations of determiner and verb object do not sound so good, but still much better than trying to put a direct object after a nominalization via a derivational morpheme such as *-ation*. Houston (1985, page 320) shows that assignment of *V-ing* forms to a discrete part-of-speech classification is less successful (in a predictive sense) than a continuum in explaining the spoken alternation between *-ing* vs. *-in'*, suggesting that "grammatical categories exist along a continuum which does not exhibit sharp boundaries between the categories."

A different, interesting example was explored by one of my graduate school classmates, Whitney Tabor. Tabor (1994) looked at the use of *kind of* and *sort of*, an example that I then used in the introductory chapter of my 1999 textbook (Manning and Schütze 1999). The nouns *kind* or *sort* can head an NP or be used as a hedging adverbial modifier:

(8) [That kind [of knife]] isn't used much.

(9) We are [kind of] hungry.

The interesting thing is that there is a path of reanalysis through ambiguous forms, such as the following pair, which suggests how one form emerged from the other.

(10) [a [kind [of dense rock]]]

(11) [a [[kind of] dense] rock]

Tabor (1994) discusses how Old English has *kind* but few or no uses of *kind of*. Beginning in Middle English, ambiguous contexts, which provide a breeding ground for the reanalysis, start to appear (the 1570 example in Example (13)), and then, later, examples that are unambiguously the hedging modifier appear (the 1830 example in Example (14)):

(12) A nette sent in to the see, and of alle kind of fishis gedrynge (*Wyclif,* 1382)

(13) Their finest and best, is a kind of course red cloth (*True Report,* 1570)

(14) I was kind of provoked at the way you came up (*Mass. Spy,* 1830)

This is history not synchrony. Presumably kids today learn the softener use of *kind/sort of* first. Did the reader notice an example of it in the quote in my first paragraph?

(15) NLP is kind of like a rabbit in the headlights of the deep learning machine (Neil Lawrence, DL workshop panel, 2015)

Whitney Tabor modeled this evolution with a small, but already deep, recurrent neural network—one with two hidden layers. He did that in 1994, taking advantage of the opportunity to work with Dave Rumelhart at Stanford.

Just recently, there has started to be some new work harnessing the power of distributed representations for modeling and explaining linguistic variation and change. Sagi, Kaufmann, and Clark (2011)—actually using the more traditional method of Latent Semantic Analysis to generate distributed word representations—show how distributed representations can capture a semantic change: the broadening and narrowing of reference over time. They look at examples such as how in Old English *deer* was any animal, whereas in Middle and Modern English it applies to one clear animal family. The words *dog* and *hound* have swapped: In Middle English, *hound* was used for any kind of canine, while now it is used for a particular sub-kind, whereas the reverse is true for *dog*.

Kulkarni et al. (2015) use neural word embeddings to model the shift in meaning of words such as *gay* over the last century (exploiting the online Google Books Ngrams corpus). At a recent ACL workshop, Kim et al. (2014) use a similar approach—using word2vec—to look at recent changes in the meaning of words. For example, in Figure 1, they show how around 2000, the meaning of the word *cell* changed rapidly from being

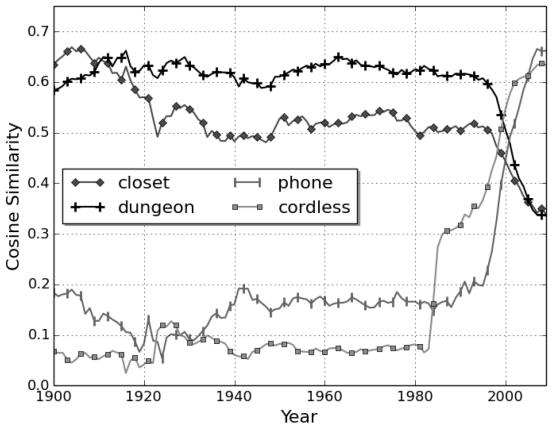

Figure 1
Trend in the meaning of *cell*, represented by showing its cosine similarity to four other words over time (where 1.0 represents maximal similarity, and 0.0 represents no similarity).

close in meaning to *closet* and *dungeon* to being close in meaning to *phone* and *cordless*. The meaning of a word in this context is the average over the meanings of all senses of a word, weighted by their frequency of use.

These more scientific uses of distributed representations and Deep Learning for modeling phenomena characterize the previous boom in neural networks. There has been a bit of a kerfuffle online lately about citing and crediting work in Deep Learning, and from that perspective, it seems to me that the two people who scarcely get mentioned any more are Dave Rumelhart and Jay McClelland. Starting from the Parallel Distributed Processing Research Group in San Diego, their research program was aimed at a clearly more scientific and cognitive study of neural networks.

Now, there are indeed some good questions about the adequacy of neural network approaches for rule-governed linguistic behavior. Old timers in our community should remember that arguing against the adequacy of neural networks for rule-governed linguistic behavior was the foundation for the rise to fame of Steve Pinker—and the foundation of the career of about six of his graduate students. It would take too much space to go through the issues here, but in the end, I think it was a productive debate. It led to a vast amount of work by Paul Smolensky on how basically categorical systems can emerge and be represented in a neural substrate (Smolensky and Legendre 2006). Indeed, Paul Smolensky arguably went too far down the rabbit hole, devoting a large part of his career to developing a new categorical model of phonology, Optimality Theory (Prince and Smolensky 2004). There is a rich body of earlier scientific work that has been neglected. It would be good to return some emphasis within NLP to cognitive and scientific investigation of language rather than almost exclusively using an engineering model of research.

Overall, I think we should feel excited and glad to live in a time when Natural Language Processing is seen as so central to both the further development of machine learning and industry application problems. The future is bright. However, I would encourage everyone to think about problems, architectures, cognitive science, and the details of human language, how it is learned, processed, and how it changes, rather than just chasing state-of-the-art numbers on a benchmark task.

Acknowledgments

This *Last Words* contribution covers part of my 2015 ACL Presidential Address. Thanks to Paola Merlo for suggesting writing it up for publication.

References

Chomsky, Noam. 1970. Remarks on nominalization. In R. Jacobs and P. Rosenbaum, editors, *Readings in English Transformational Grammar*. Ginn, Waltham, MA, pages 184–221.

Houston, Ann Celeste. 1985. *Continuity and Change in English Morphology: The Variable (ing)*. Ph.D. thesis, University of Pennsylvania.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International World Wide Web Conference (WWW 2015)*, pages 625–635, Florence.

Luong, Minh-Thang, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Curran Associates, Inc., pages 3111–3119.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543, Doha.

Prince, Alan and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell, Oxford.

Ross, John R. 1972. The category squish: Endstation Hauptwort. In *Papers from the Eighth Regional Meeting*, pages 316–328, Chicago.

Rumelhart, David E. and Jay L. McClelland, editors. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. MIT Press, Cambridge, MA.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In Kathryn Allen and Justyna Robinson, editors, *Current Methods in Historical Semantics*. De Gruyter Mouton, Berlin, pages 161–183.

Smolensky, Paul and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, volume 1. MIT Press, Cambridge, MA.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Curran Associates, Inc., pages 3104–3112.

Tabor, Whitney. 1994. *Syntactic Innovation: A Connectionist Model*. Ph.D. thesis, Stanford.