Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Natural Language Processing

This input paper was prepared by Tim Baldwin and Karin Verspoor

Suggested Citation

Baldwin, T and Verspoor, K (2018). Natural Language Processing. Input paper for the Horizon Scanning Project “The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing” on behalf of the Australian Council of Learned Academies, www.acola.org.

AI AND NATURAL LANGUAGE PROCESSING

Timothy Baldwin (The University of Melbourne) Karin Verspoor (The University of Melbourne)

Background

Natural Language Processing ("NLP") encompasses all technologies related to the analysis and generation of text- based natural language, with prominent applications including machine translation (e.g. Google Translate), dialogue systems (e.g. the back end systems that underlie Apple's Siri and Amazon's Alexa), and automatic question answering (e.g. IBM Watson). NLP is a core pillar of Artificial Intelligence; this is a natural corollary of the critical importance of language to the evolution and advancement of the human race, as the basis of interpersonal communication, developing and disseminating scientific and social theory, structuring and enforcing legal agreements, and archiving our history. It has been argued to shape thought and to provide the foundation for social identity. In short, language is a fundamental aspect of being human.

While written language is a relatively recent development, it has become a cornerstone of our world. The advent of the Internet has enabled rapid communication at massive scale, resulting in an unprecedented amount of language being produced, shared, and recorded in electronic textual form. Some analyses indicate that over 3.7 billion people use the internet, executing 5 billion searches per day, close to 500k tweets and 300k status updates every minute [Marr, 2018]. This leaves aside the spoken language available in the millions of videos that are produced. All of this language data demands analysis through automated means – how else can we sift through all of the information conveyed by that text? – and represents both significant opportunities and significant challenges.

NLP has matured substantially in the last decade, to a point where it is strongly contributing to automate/streamline various tasks, in terms of accuracy and tractability in production systems. Some of the key elements of this maturation have been: better language models (meaning more reliable/fluent natural language outputs), a move away from sparse word-based representations to dense representations in latent semantic spaces (associated with better semantic generalisation, and richer cross-lingual mappings), a move towards character- rather than word-based models (leading to better handling of rare, misspelled, and otherwise low-frequency words), improvements in large-scale model training/integration (making it possible to "pre-train" models over large-scale datasets, and compose component models into a single trainable model), and substantial improvements in the ability to train models over multimodal inputs (e.g. text and images, vastly improving the accuracy of models at tasks such as image captioning and question answering over images). Much of this has been driven by "deep learning", which pervades modern-day NLP.

The Coming Decade for NLP

Given this background, what advances in NLP capabilities can we expect to see in the next decade, and what appli- cations will they be tied to?

Advances

Development of models which can justify their outputs to humans

As the models used in NLP become more and more complex — in terms of the raw number of parameters, and the interactions between those parameters — it becomes increasingly difficult to trace back through the model to explain why it produced the output it did. This has implications from pure scientific advancement (i.e. in order to improve a model, we ideally need to know what it has learned and not learned from a given training dataset) to user trust (e.g. in a clinical setting, a doctor needs to be able to scrutinise the basis of a diagnosis or treatment recommendation to be able to treat a patient) and legal compliance (e.g. the "right to explanation" under the terms of the GDPR). Explainability will become increasingly critical

as AI is deployed in human-facing settings in ever more complex settings, with the most naturalistic user interface being language-based (e.g. in the form of a dialogue system, to interact with the model to understand why the system performed in the way it did). As such, NLP will have a critical role to play across the full spectrum of AI in terms of explainability.

NLP with world knowledge

While there has been increased use of knowledge bases such as YAGO and Freebase in NLP (which contain entity graphs with links encoding the relationships between entities, such as Bill Gates being the founder of Microsoft), there has been almost no progress on the lack of world knowledge in NLP systems. An example of the need for world knowledge can be seen in “Winograd Schemata” [Levesque et al., 2011] such as: The large ball crashed right through the table because it was made of steel, where the task is to determine which of the ball and table were made of steel (i.e. de-reference it), and examples are carefully constructed such that superficial artefacts in the examples do not give away the answer. In some specific domains, such as biomedicine, the possibility of a closed-world assumption

— where everything that is known is formally introduced and catalogued — and availability of structured knowledge resources has enabled some progress towards deeper integration of reasoning over knowledge and NLP. In the next decade, we anticipate the development of large-scale general-purpose knowledge bases that go beyond “entities” (e.g. Bill Gates or Medicare) with structured data fields (e.g. date-of-birth or country-of-origin) to include both more intangible notions such as time and complex phenomena such as the wind. Equally, we expect such knowledge bases to capture semi-structured information ranging from physical properties/processes (e.g. how is wind generated) to cultural manifestations (e.g. how is the wind represented in impressionist art) or socio-economic implications (e.g. the cost to Australian agriculture of wind-based erosion). NLP will play a critical part in the semi-automatic construction of such a knowledge base, and benefit directly from harnessing that knowledge, with substantial technical challenges in how to make efficient use of data of this scale and complexity.

Cross-domain and cross-task robustness

Most of the advances in NLP over the last decade have been achieved over specific tasks/datasets, driven by ever- larger datasets, fueling ever more data-hungry models, and based on tailored solutions to the particular task. Unlike humans, the ability of an NLP system to perform one language task well generally does not indicate general language facility: when highly-successful models for one task are applied to a closely related task (e.g. a question-answering model is applied in the context of a tutoring system) or even the same task in a different domain (e.g. a question- answering model trained over Wikipedia data being applied over biomedical literature), the drop in results is generally huge. We expect there to be significant advances in general-purpose language processing through cross-training across multiple tasks and explicit domain debiasing, such that off-the-shelf system components — possibly in com- bination with domain-specific plug-and-play knowledge bases, and data-efficient methods for tuning a method to a particular task — can be applied to novel tasks/domains, with reasonable expectation of competitive performance.

Algorithmic equity without compromising overall system accuracy

In the last 12 months, there have been a number of key publications exposing the confronting reality that when naively trained, NLP models not only mimic but tend to accentuate bias in the underlying datasets , leading to systems which work better for users who are overrepresented in the training data (in general, US English-speaking white, middle-aged males), leading to inherent inequities in the ability of different populations to tap into the benefits of AI. Similarly, models tend to learn more subtle but equally troubling dataset biases, such as that doctors are male and nurses are female, and reflect these in system outputs (e.g. in pronoun choice in document summaries and translations). To alleviate such biases, we generally need to have explicit knowledge of the existence of the bias, and then use explicit training data to mitigate that bias. This is problematic, however, as we have no way of documenting all of the myriad dimensions of bias in our training datasets and the relative social impact of each of them, and in order to remove bias with respect to highly sensitive issues such as sexuality, e.g., we need training users to expose this

data about themselves to be able to measure and ultimately remove any associated bias. There are also potentially legal compliance aspects to the problem, e.g. in automatically screening job applicants, in most cases it would be illegal to discriminate against on the basis of gender, race, or disability. Simply removing explicit information about gender, e.g., generally doesn't solve the problem, as models may be able to implicitly infer this from other personal traits which are relevant to suitability of the candidate to the role such as education history (e.g. if a candidate attended a single-sex school), or the combination of citizenship and compulsory military service. Even more challenging is determining when it is legal to use such personal information, e.g. in the case of a male applying for an explicitly female-only position (with appropriate legal exclusion), or excluding a candidate with a physical disability that would prevent them from fulfilling the duties of a position with a mandatory travel requirement.

Better context modelling

The vast majority of NLP systems still operate at the sentence level.[1] That is, the first thing that happens when processing a document is to partition it up into its component sentences, and process them one at a time, independent of one another. This has obvious disadvantages in terms of consistency (e.g. translating the same word in a given document inconsistently across different sentences of that document), and document flow/cohesion (e.g. ignoring cross-sentential pronoun references, or using inappropriate discourse connectives). We expect to see large advances in the modelling of context, beyond simple document context to include social context (e.g. personalising the translation based on the identity of the author and their social network, the intended audience for the translation, or a particular viewpoint on the content) or author demographics (e.g. personalising the translation of a document or the output of a chatbot to a particular persona, in terms of age, gender, language background, etc.). In order to achieve this, we will need more expressive models, much more fine-grained control over system outputs, as well as challenge datasets with appropriate richness and variety to be able to model such subtleties. One particular area with considerable scope for improvement is multi-turn dialogue systems in knowledge-rich, dynamic domains, e.g. an interactive tutoring system to debate/understand complex societal issues from different viewpoints, which is able to comprehend and recall the full interaction history with a user; or a dialogue system which monitors/manages a complex environment and is able to provide rich diagnostics with historical context (e.g. *the light on the back patio is playing up, stirring up the dogs and likely causing the neighbours to complain again*).

Multimodal processing

When humans learn language, they do so in a rich, situated context using all of their senses with a myriad of feedback mechanisms. There has been a growing body of work on multimodal AI — most notably combining text and image analysis — but equally results to suggest that many of the early successes have been data/modelling "accidents" rather than methodological breakthroughs. Learning from these findings, we expect there to be real breakthroughs which integrate different modalities of input, with virtuous interactions between the component modalities leading to standalone gains in NLP performance, richer capture of world knowledge, and enhanced model interpretability.

Progress on task-oriented discourse processing

There are many environments where hands-free, language-based, interaction with an intelligent agent would enable more effective decision-making. We see this already in limited applications such as automobile navigation systems, which support the sort of "single-turn" dialogues, with a clear objective. We expect the sophistication of such systems to increase dramatically, utilising more detailed modeling of specific task contexts to enable de-referencing of entity mentions and tracking of relationships between entities.

In addition, we foresee increased coupling of question answering applications with conversational agents, e.g. for customer service "bots" on websites that support more flexible interaction with customers, to direct them to resources on the website or to provide a direct response to a query. A key challenge will be

accommodating broad information seeking questions that are much more open-ended and complex than the typical web search query issued today.

2.2 Applications

Several key application areas are emerging that we can expect to attract substantial attention.

References

1. D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan,
2. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain. The science of fake news. Science, 359(6380):1094–1096, 2018. ISSN 0036-8075. doi: 10.1126/science. aao2998. URL http://science.sciencemag.org/content/359/6380/1094.
3. J. Levesque, E. Davis, and L. Morgenstern. The Winograd Schema Challenge. In Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, pages 552–561, Rome, Italy, 2011.1With notable instances of document-level tasks such as document summarisation.
4. B. Marr. How much data do we create every day? the mind-blowing stats every- one should read. https://www.forbes.com/sites/bernardmarr/2018/05/21/ how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/ #1e1cc9d260ba, 2018. Accessed: 2018-08-14.
5. F. Menczer and A. Flammini. Observatory on social media. https://truthy.indiana.edu/, 2011-2018. Accessed: 2018-08-14.