On the Challenges of Translating NLP Research into Commercial Products

Daniel Dahlmeier
SAP Innovation Center Singapore
d.dahlmeier@sap.com

Abstract

This paper highlights challenges in industrial research related to translating research in natural language processing into commercial products. While the interest in natural language processing from industry is significant, the transfer of research to commercial products is non-trivial and its challenges are often unknown to or underestimated by many researchers. I discuss current obstacles and provide suggestions for increasing the chances for translating research to commercial success based on my experience in industrial research.

1 Introduction

Natural language processing (NLP) has made significant progress over the last two decades, in particular due to the success of data-driven machine learning methods. Recently, deep learning has led to another wave of remarkable improvements in NLP and other areas of machine learning and artificial intelligence (AI). Not surprisingly, many industry players are investing heavily in machine learning and AI to create new products and services (MIT Technology Review, 2016).

However, translating research into a successful product has its own challenges. Traditionally, technology transfer is often assumed to happen in a linear transition from pure research to applied research to commercialization (Stokes, 1997). The model assumes that the discoveries from researchers will naturally be picked up by engineers and industry players who will use it to build new products. In reality, the transfer from research to commercial products is considerably more complex and far from guaranteed. In fact, many research projects fail to successfully transfer their discoveries to commercial products.

In this position paper, I highlight some of the reasons why it is so difficult to translate NLP research into successful products. This paper does not contain any new algorithms, experiments, or results. Instead, it seeks to share my experience working at the intersection of academic research and industry with the aim to stimulate a discussion how technology transfer of NLP research can be improved. I want to emphasize upfront that the paper is not arguing that all NLP researchers should focus their efforts on building commercial products nor does every new product require a research breakthrough to be successful. The paper's aim is rather to discuss how we can improve *use-inspired basic research* that satisfies both the desire for fundamental understanding and considerations of use, sometimes referred to as *Pasteur's quadrant* (Stokes, 1997).

The contributions of this paper are twofold. First, I highlight common obstacles in the path of transferring research into commercial products. Second, I offer suggestions for increasing the chances of success based on my experience at SAP, the world's largest enterprise software company.

2 Challenges to Innovation

This section highlights challenges in NLP research that make it difficult to translate the results into impactful innovation.

2.1 Lack of Value Focus

The first step to creating a successful product is understanding your customers. That is why many methodologies for creating new products or business models start with a *user persona* and how to create value for the user (Ries, 2011; Osterwalder et al., 2014). Similarly, to conduct research with practical impact, it is worthwhile to consider what potential applications the research could enable

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 92–96
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2015

and what the *value proposition* for a potential user might be. The value proposition is closely linked to the user persona and the tasks that she tries to solve in her daily life (Christensen and Raynor, 2013). Thus, choosing the right research task is important when aiming for impactful research. It is instructive that NLP tasks which solve practical problems, like machine translation or sentiment analysis, have seen significant adoption in commercial applications. But many applications that are requested by industry are still beyond the capabilities of current NLP research, for example chatbots that can respond to arbitrary user questions.

It is also important for researchers to understand that the priorities in industry are different from priorities in academic research. In academic research, the priorities are to create contributions to the body of knowledge in the field, e.g., defining a new task, a novel, elegant model, or a new state-of-the-art benchmark result. In industry the priorities are creating innovative products that delight users and create new revenue streams. To have the best of both worlds, researchers should occasionally take a step back and consider what value proposition their work has for people outside the NLP community.

2.2 Lack of Reproducibility

Reproducible research is one of the pillars of the scientific method and thus important to good research work in general. But the ability to reproduce a model is also a prerequisite to incorporating it into a product. As NLP models often depend on a complex set of parameters and pre-processing steps which cannot always be explained in all detail in a paper, it is often hard to reproduce other's results. The author himself has his own experience trying to (unsuccessfully) reproduce published results. As problems to reproduce research are seldom reported (but see (Bikel, 2004) for an exception), it is also hard for researchers to find information on how to improve their implementation when they struggle to re-produce published results.

2.3 Lack of (Domain) Data

Data is the fuel that powers machine learning and most of NLP research. While the "big data" revolution has given us access to large quantities of text data from some domains, for many industry problems there is no or very limited data available to conduct research on. For example, in my group we have been working on text classification for customer service tickets. While there are many datasets available for text classification, these are primarily from newswire or online reviews. For customer service, there is no public dataset to compare to. Due to the confidential nature of the data and data privacy concerns, companies who have such data cannot easily release it for research purposes. Some companies host shared tasks or data science competitions in which they make data available, for example on Kaggle[1], but access to data remains one of the biggest obstacles for researcher who want to work on industry problems.

Even when there is data available from public sources, e.g., from the web, using the data for commercial purposes can be tricky from a legal standpoint. Crawling data from web (or using corpora created by others in this manner) might be acceptable for research purposes, but when building a commercial product the exact license, copyright, etc. of every data source needs to be checked. The same holds for publicly available NLP models derived from such data.

For everyone who believes that working in industry solves all data problem, I note here that working with real data sets has its own challenges. Real data sets are often small, noisy, scrambled, or otherwise incomplete, making it hard to achieve good results. To effectively use the data, researchers also have to understand the data schema and the business process behind the data. This can be challenging without and in-depth domain knowledge.

2.4 Overemphasis on Test Scores

The empirical evaluation of statistical methods on common benchmarks has without a doubt revolutionized NLP (Johnson, 2009). However, sometimes the score on the test set is taken as the *only* factor that determines the success of a piece of research. For practical applications, the test score on a benchmark dataset is only one criteria among many when it comes to choosing an algorithm for practical use. Other factors include the time and costs required to implement the method, the computational resources required, speed and performance, the ease of integration, support for multi-lingual input, the ability to adapt and customize the method, the ability to incorporate prior knowledge, and the ability to interpret and explain

[1] https://www.kaggle.com

the model. For example, in our text categorization work, we encountered the requirement to accommodate changes in the the output classes, i.e., adding, merging, splitting, and removing classes, without re-training the model from scratch. These factors are currently underrepresented in NLP research.

2.5 Difficulty of Adoption

No matter how good an NLP model is, it cannot have practical impact if it is never implemented. But in any application, the NLP model will only be one component in a larger software system. How easily the NLP component can work together with the remaining components is important for the ease of adoption of the method into productive applications. Unlike rule-based methods, statistical NLP models often require expensive collection and labeling of data, data pre-processing, model (re-)training, parameter tuning, and monitoring of the model to avoid model staleness. This makes it harder to adopt statistical models in practical applications (Chiticariu et al., 2013).

2.6 Timelines

The time horizon within which stakeholders expect results is generally shorter in industry projects. While research grants typically run for three to five years, industry research is under pressure to deliver tangible outcomes in less than two years. For projects with actual customers and proof of concepts, timelines are usually not longer than a few months. This results in the following chicken and egg problem: it is difficult to produce groundbreaking research within a short time frame but long investments into research are hard to justify if the value the research will ultimately produce is not clear. That is why academic research is generally better equipped to focus on fundamental research questions. Fundamental research does not exclude practical usage but incremental research that fine-tunes every aspect of the implementation of an NLP model is often better done in industry labs.

3 Bridging the Gap

In this section, I offer some suggestions about how the disconnect between NLP research and commercial products can be reduced.

3.1 A "Recipe" for Qualifying a Research Problem

The following approach describes the criteria that we typically apply in our team when we evaluate new machine learning use cases, including NLP use cases.

First, we make sure we understand the "job to be done": what is the business problem, who is the potential user and what problem are we trying to solve? Once we have understood the task, a first question to ask is whether the task actually requires NLP. Is the data volume so high that automation is needed? Would it be easier or cheaper to solve the task manually? Can the task be solved via simple rules? Typically, tasks with high data volume and complex or ambiguous rules are good candidates for NLP.

To ensure that the use cases we work on have practical relevance, we include stakeholders from the lines of business and industry units in the company in any new project right from the beginning and gather feedback from actual customers.

Once we believe that NLP is required, we try to formulate the problem as a machine learning task. The simple template *given X, predict Y* together with the question *what are the inputs and what are the outputs?* helps significantly to get from a vague idea to a concrete task formulation. At this stage, we can often already map the problem to a standard NLP task, e.g., text classification, sequence tagging, or sequence-to-sequence learning.

Next, we establish whether data is available. If real data is not available easily, can we work with publicly available proxy data? For example for learning to classify customer service tickets, we can start with text classification on public datasets. If it is unlikely that data will be available in the foreseeable future, we do not proceed with a use case.

Next, we make a best guess whether the problem can be solved with the current state of the art in NLP. Is there an intuitive regularity in the data which we believe a statistical model could pick up? Can we represent the input via meaningful features? Do we have a way to measure the success of the method with a well-defined metric?

Finally, we determine the right approach to execute the use case. If it is a hard problem which needs at least a few more years of research before it becomes useful, we would most likely decide on a research project. We fund external

research projects at top universities around the world, where we provide the research problem and the data and let others try to crack the tough problems. We also sponsor Ph.D. students who are working at SAP during their studies.

If we think that the use case has a strong business case and the technology is mature enough, we will move it to building a proof of concept, and ultimately a commercial product. While this "recipe" for qualifying an NLP use case is simple and common sense, we have found it helpful in prioritizing use cases.

Researchers in academia might not have access to a business unit to provide feedback on research ideas but many funding bodies are trying to encourage increased collaborations between industry and academia. The European Union, for example, has specifically funded an initiative, LT-innovate[2] to encourage commercial exploitation of NLP research.

3.2 Engineering Approach to NLP

I believe that a more rigorous application of (software) engineering principles and tools can greatly increase the odds of having practical impact with NLP research.

To address the problem of reproducibility, I suggest the following. First, the community should be more stringent about reproducibility. In some research communities, for example databases, the criteria for reproducible research are a lot stricter. If the results are not reproducible, the results are generally not considered valid. However, the large number of parameters and implementation details in NLP systems makes it hard to exactly reproduce published results based on the paper alone. Therefore, we should encourage the dissemination of results through software tools that make code reproducible. To reproduce the results in a paper, we essentially need the code, the data, and the parameters of the experimental run that produced the results of the experiment. Fortunately, the open source community has created great tools that make this possible. First, social code repository platforms, such as GitHub[3], make it easy to share code and data. In fact, the ease of sharing and contributing to code has arguably accelerated the progress in machine learning significantly. Second, interactive computational environments, such as Jupyter notebooks[4], that tie together data, code, and documentation, allow for reproducible results that can easily be shared and published. Finally, software containers, such as Docker[5], allow light-weight virtualization that pulls in all software dependencies and allow the same code to run in a reliable and reproducible manner. If a Jupyter notebook or Dockerfile is published with the paper, it should be easier for other researchers to reproduce results and integrate them into larger systems. Projects like CodaLab[6] try to build online platforms for reproducible research with similar goals.

On the problem of data availability, there is already a considerable amount of work in the area of building NLP models in low-resource environments (see for example (Duong et al., 2014; Garrette and Baldridge, 2013; Wang et al., 2015)) which deals with limited data availability. Techniques like domain adaptation, semi-supervised learning and transfer learning (Pan and Yang, 2010) are extremely relevant to address the problem of data availability for industry applications. Finally, recent work on learning models from private data (Papernot et al., 2016) and federated learning across many devices (McMahan et al., 2016) appear to be promising directions for practical NLP engineering research.

3.3 Industry Papers

I believe that there is an opportunity to increase the exchange between industry and the research community by establishing an industry paper submission format, potentially with its own industry track at NLP conferences. Such a track could offer a venue to discuss practical challenges in building large-scale NLP systems and deploying NLP models in production settings, such as scalability, trade-offs between accuracy and computational costs, robustness, data quality, etc. This would help to counter-balance the overemphasis on test scores in pure research papers and aid the adoption of research in industry applications. Industry tracks are common in other communities and have strong participation from industry players there.

[2] http://www.lt-innovate.org
[3] https://github.com/
[4] http://jupyter.org/
[5] https://www.docker.com/
[6] http://codalab.org/

4 Related Work

Wagstaff (2012) argues for making machine learning research more relevant. He laments a hyperfocus on UCI benchmark datasets and abstract metrics. Spector *et al.* (2012) present Google's hybrid approach to research, which tries to avoid separation between research and engineering. Recently, several groups at Google have published papers on practical challenges in deploying machine learning in production (Sculley et al., 2014; McMahan et al., 2013; Breck et al., 2016). Belz (2009) discusses the practical applications of NLP research. Mani (2011) gives suggestions for improving the review process. None of the works provides a detailed discussion on the difficulties in bringing NLP research to commercial products – the main contribution of this paper.

5 Conclusion

I have highlighted difficulties that exist for researchers who try to bring NLP research into commercially products and offered suggestions for improving the odds of commercial success. I hope that my experience can stimulate creative thought and a healthy discussion in the NLP community.

References

Anja Belz. 2009. That's nice... what can you do with it? *Computational Linguistics* 35(1).

Daniel Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics* 30(4).

Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D. Sculley. 2016. Whats your ML test score? A rubric for ML production systems. In *Proceedings of Reliable Machine Learning in the Wild - NIPS 2016 Workshop (2016)*.

Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of EMNLP*.

Clayton M. Christensen and Michael E. Raynor. 2013. *The Innovator's Solution: Creating and Sustaining Successful Srowth*. Harvard Business Review Press.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? A case study of multilingual POS tagging for resource-poor languages. In *Proceedings of EMNLP*.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of HLT-NAACL*.

Mark Johnson. 2009. How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*.

Inderjeet Mani. 2011. Improving our reviewing processes. *Computational Linguistics* 37(1).

H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. 2013. Ad click prediction: a view from the trenches. In *Proceedings of ACM SIGKDD*.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agera y Arcas. 2016. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of AISTATS*.

MIT Technology Review. 2016. AI Takes Off. https://www.technologyreview.com/business-report/ai-takes-off/. Online; accessed 22 April 2017.

Alexander Osterwalder, Yves Pigneur, Gregory Bernarda, and Alan Smith. 2014. *Value proposition design: How to create products and services customers want*. John Wiley & Sons.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10).

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2016. Semi-supervised knowledge transfer for deep learning from private training data.

Eric Ries. 2011. *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Crown Business.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. 2014. Machine learning: The high interest credit card of technical debt. In *Proceedings of SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*.

Alfred Spector, Peter Norvig, and Slav Petrov. 2012. Google's hybrid approach to research. *Communications of the ACM* 55(7).

Donald E. Stokes. 1997. *Pasteur's quadrant: Basic Science and Technological Innovation*. Brookings Institution Press.

Kiri Wagstaff. 2012. Machine learning that matters. In *Proceedings of ICML*.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of ACL*.