JUNE 30, 2016

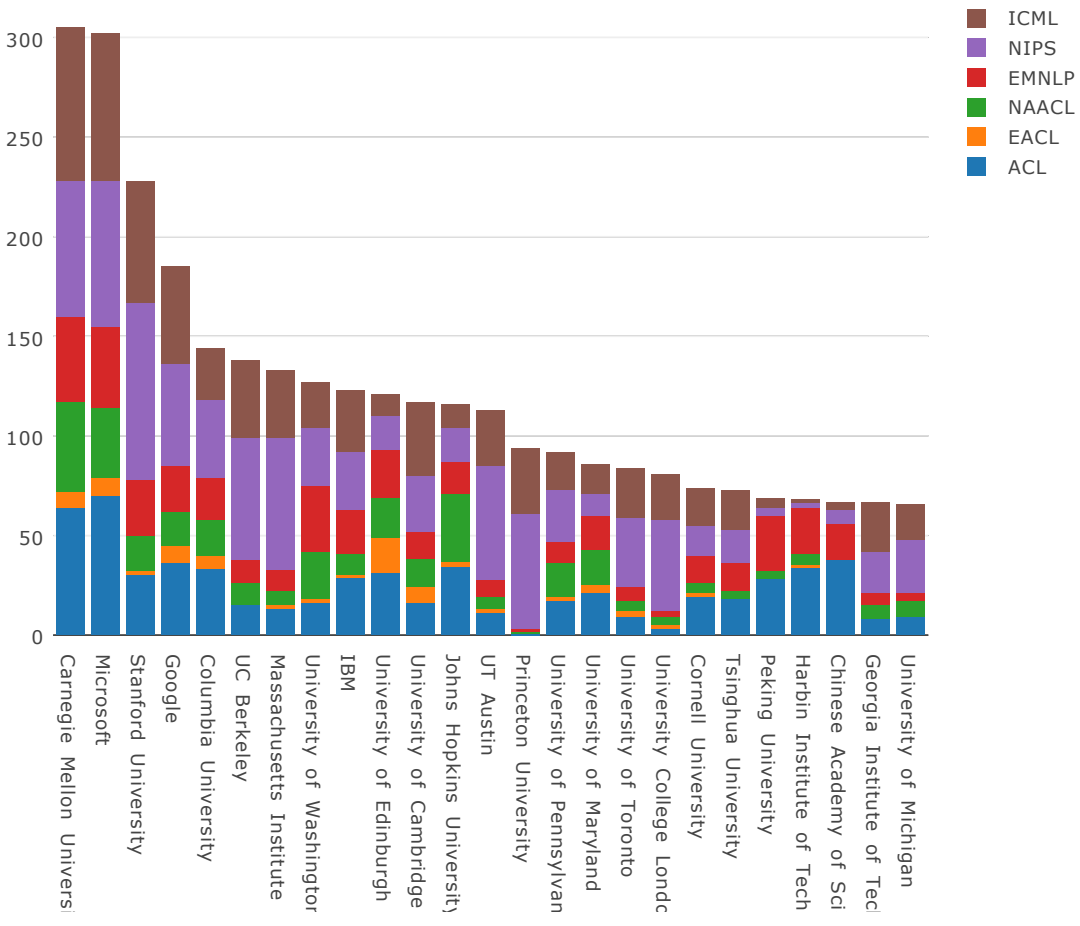
Analysing NLP publication patterns

Recently, I got curious about finding out how much different institutions publish in my area. Does Google publish more than Microsoft? Which university has the strongest publication record in NLP? And are there any interesting trends that can be seen in the recent years? Quantity does not necessarily equal quality, but the number of publications is still a reasonable indicator of general activity in the field, how big the research group is, and how outward-facing are the research projects.

My approach was to crawl papers from the 6 biggest conferences that are relevant to my research: ACL, EACL, NAACL, EMNLP, NIPS, ICML. The first 4 focus on NLP applications regardless of methods, and the latter 2 on machine learning algorithms regardless of tasks. The time window was restricted to 2012-2016, as I'm more interested in current publications.

Luckily, all these conferences have nice webpages listing all the papers published there. ACL Anthology (http://aclweb.org/anthology/) contains records for ACL, EACL, NAACL and EMNLP, NIPS (https://papers.nips.cc/) has a separate webpage for papers, and ICML proceedings are on the JMLR (http://jmlr.org/proceedings/papers/v28/) website (except for ICML12 (http://icml.cc/2012/papers/) which are on the conference website). I wrote python scripts that crawled all the papers from these conferences, extracting author names and organisations. While authors can be crawled directly from the websites, in order to find the organisation names I had to parse the pdfs into text and extract anything that looked like a university or company name in the first 30 lines of on the paper. I wrote a bunch of manual patterns to map names to canonical versions ("UCL" to "University College London" and "Google Inc" to "Google"), although it is likely that I still missed some edge cases.

Below is the graph of top 25 organisations and the conferences where they publish.

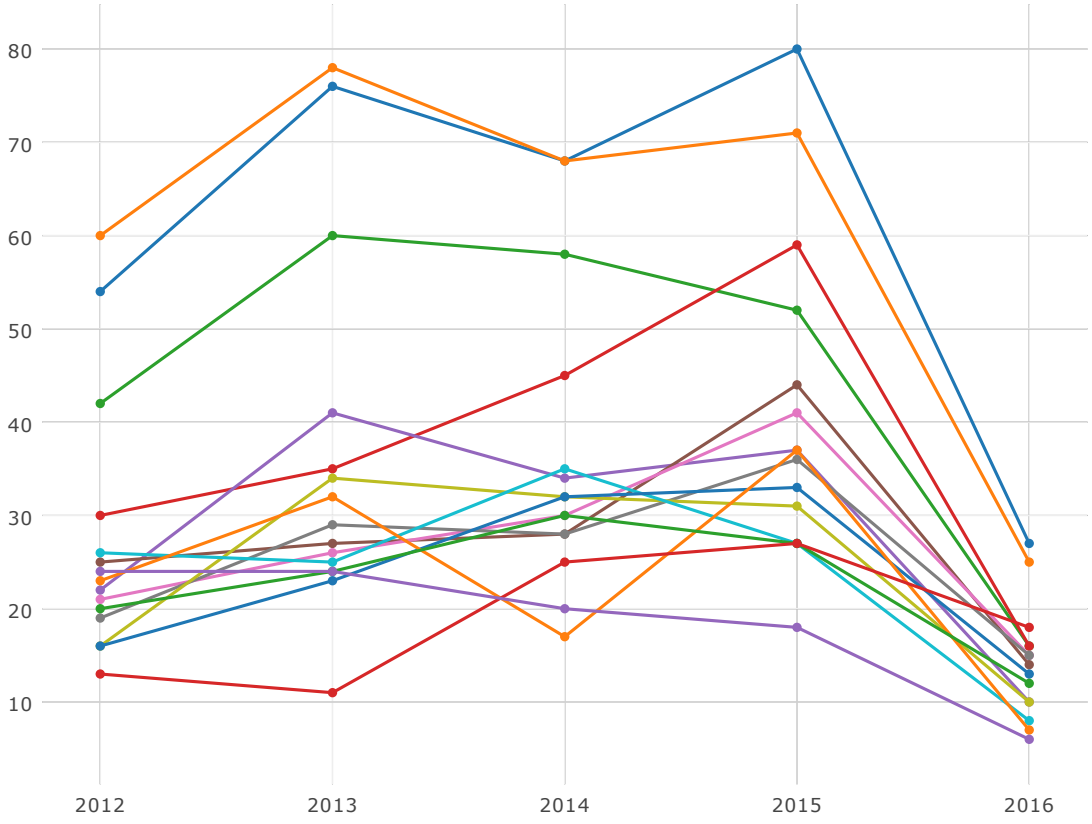


ity of Technology ia on hnology ences nology

CMU comes out as the most prolific publisher with 305 papers. A close second is Microsoft with 302 publications, also leading in the industry category. I was somewhat surprised to find that Microsoft publishes so much, almost twice as many papers compared to Google, especially as Google seems to get much more publicity with their research. Stanford is also among the top 3 organisations that publish substantially more than others. Edinburgh and Cambridge represent the UK camp with 121 and 117 papers respectively.

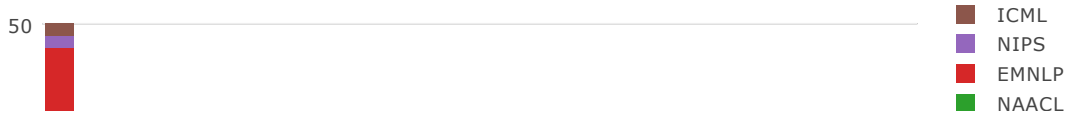
When we look at the distribution of conferences, Princeton and UCL stand out as having very little NLP-specific research, with nearly all of their papers in ICML and NIPS. Stanford, Berkeley and MIT also seem to focus more on machine learning algorithms. In contrast, Edinburgh, Johns Hopkins and University of Maryland have most of their publications on NLP-related conferences. CMU, Microsoft and Columbia are the most balanced among the top publishers, with roughly 50:50 division between NLP and ML.

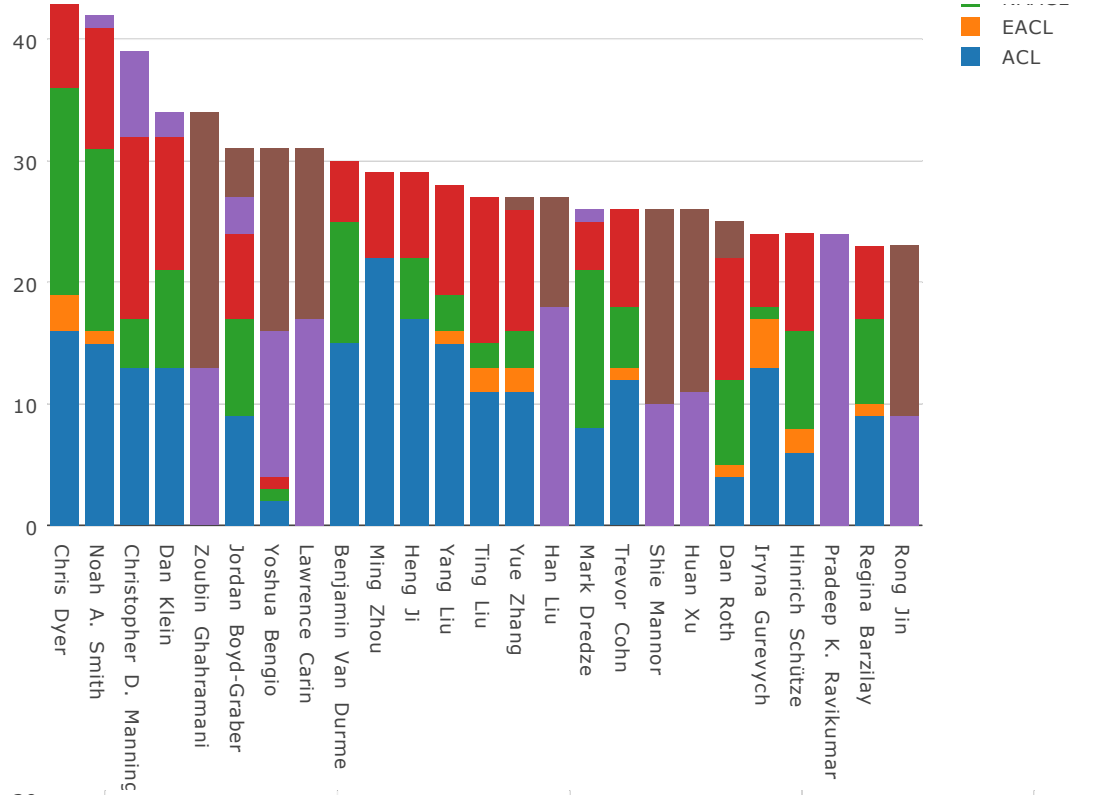
We can also plot the number of publications per year, focusing on the top 15 institutions.

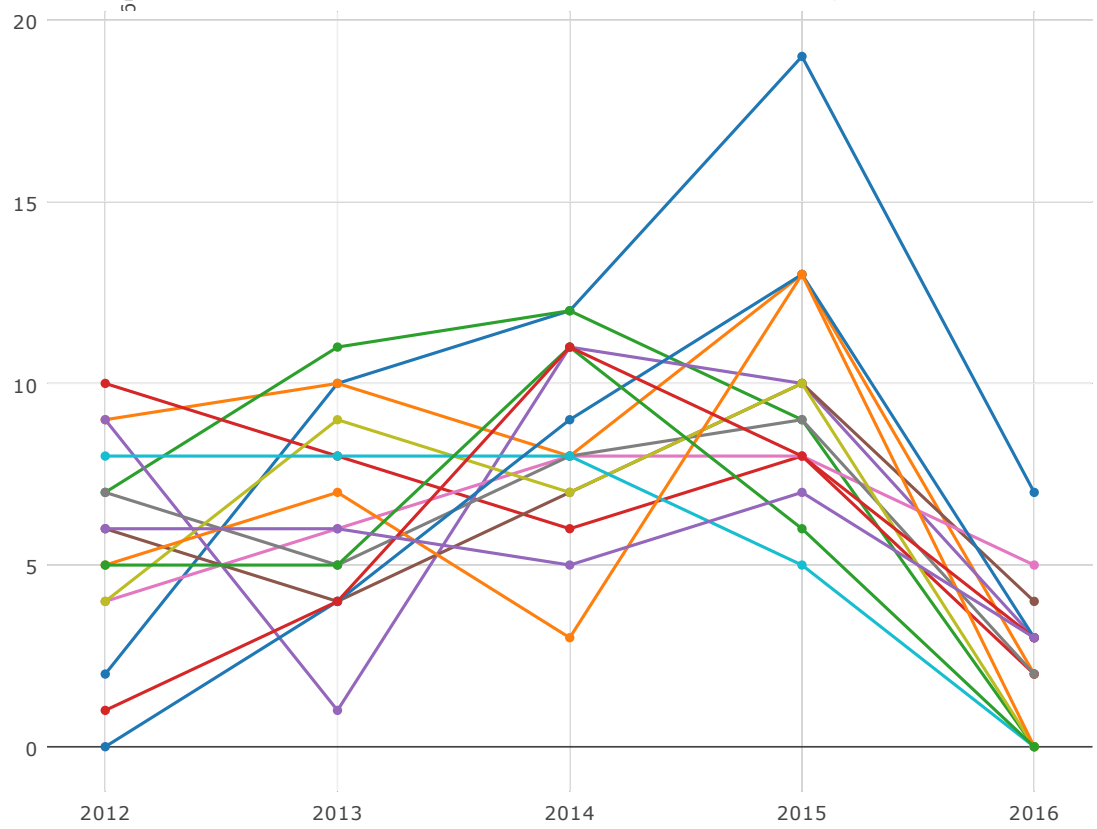


Carnegie Mellon has a very good track record, but has only just recently overtaken Microsoft as the top publisher. Google, MIT, Berkeley, Cambridge and Princeton have also stepped up their publishing game, showing upward trends in the recent years. The sudden drop for 2016 is due to incomplete data – at the time of writing, ACL and NIPS papers for this year are not available yet.

Now let's look at the same graphs but for individual authors.







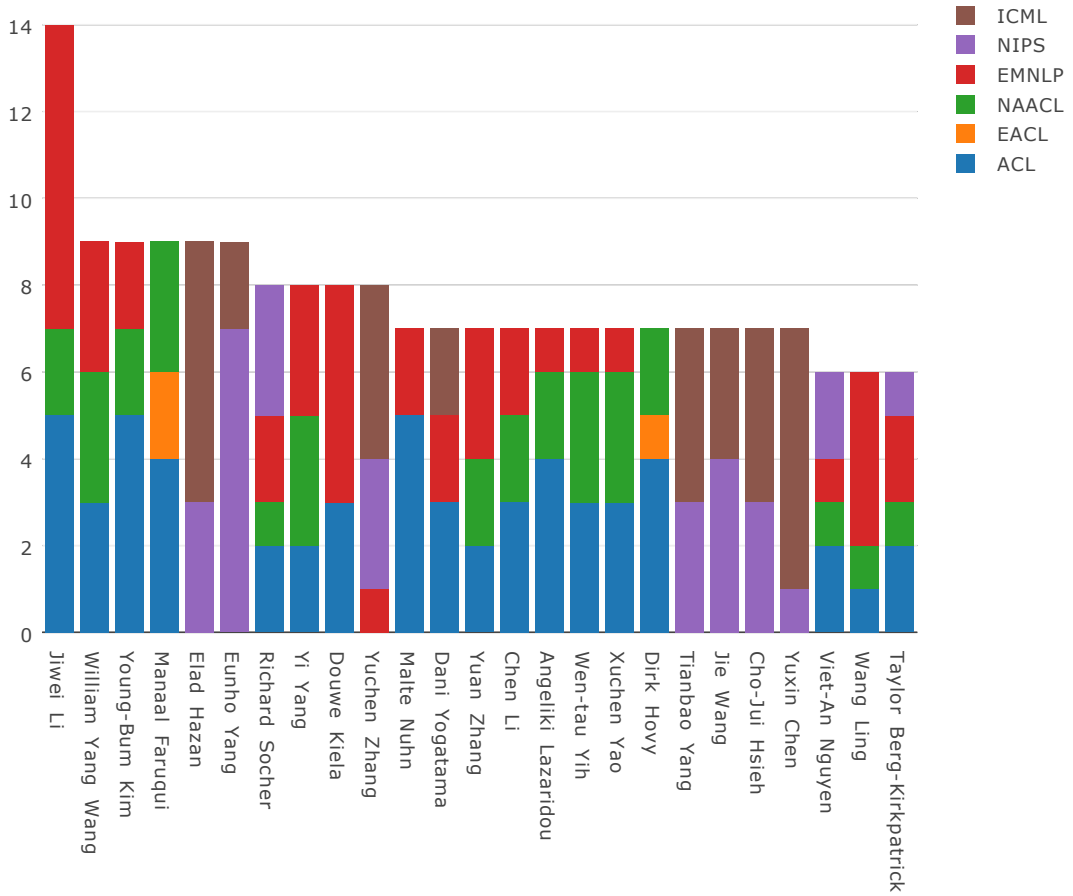
Chris Dyer comes out on top with 50 papers. This result is even more impressive given that he started with just 2 papers in 2012, then rocketing to the top by quite a margin in 2015. Almost all of his papers are in NLP conferences, with only 1 paper each for NIPS and ICML. Noah Smith, Chris Manning and Dan Klein rank 2nd-4th, with more stable publishing records, but also focusing mainly on NLP conferences. In contrast, Zoubin Ghahramani, Yoshua Bengio and Lawrence Carin are focused mostly on machine learning algorithms.

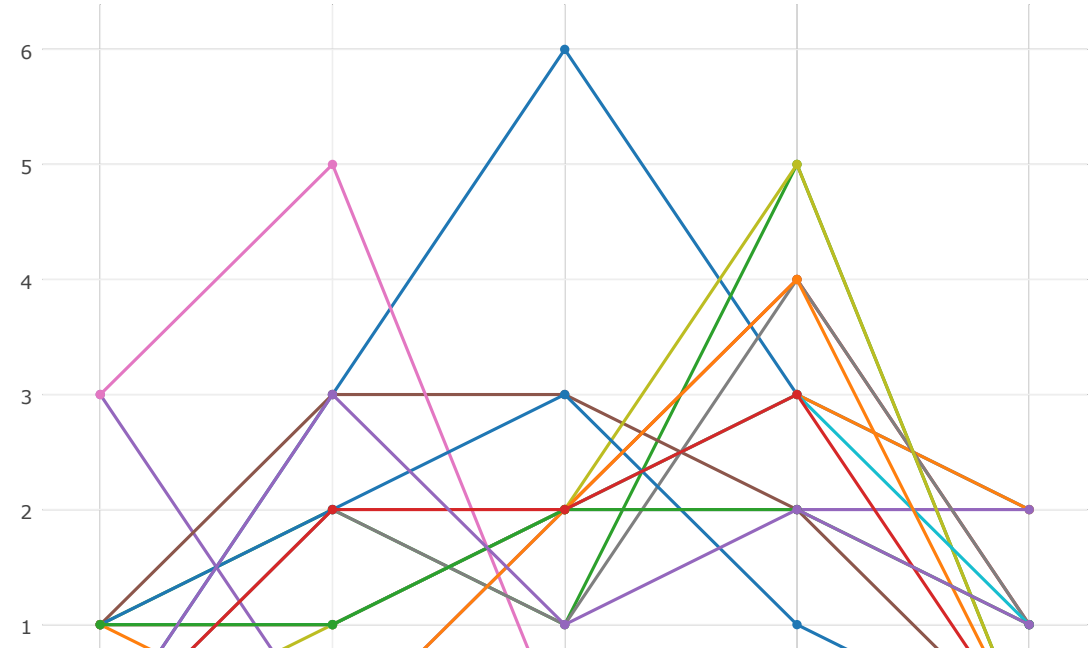
There seems to be a clear separation between the two research communities, with researchers specialising to publishing either in NLP or ML. This seems somewhat unexpected, especially

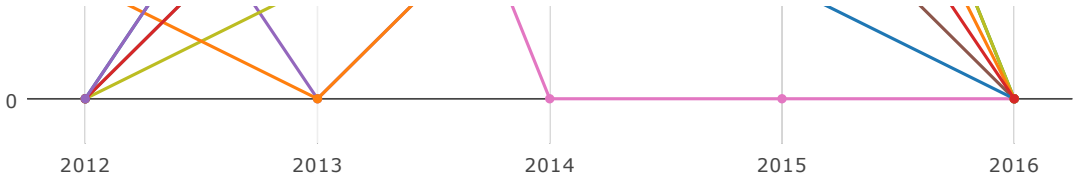
considering the widespread trend of publishing novel neural network architectures for NLP tasks. Both fields would probably benefit from slightly tighter integration in the future.

I hope this little analysis was interesting to fellow researchers. I'm happy to post an update some time in the future, to see how things have changed. In the meantime, let me know if you find any bugs in the statistics.

Update: As requested, I've also added the statistics for first authors with highest publication counts. Jiwei Li (http://web.stanford.edu/~jiweil/) from Stanford towers above others with 14 publications. William Yang Wang (https://www.cs.cmu.edu/~yww/) (CMU), Young-Bum Kim (https://www.microsoft.com/en-us/research/people/ybkim/) (Microsoft), Manaal Faruqui (http://www.cs.cmu.edu/~mfaruqui/) (CMU), Elad Hazan (http://www.cs.princeton.edu/~ehazan/) (Princeton), and Eunho Yang (IBM) have all managed an impressive 9 first-author publications.







Update 2: Added a fix for Jordan Boyd-Graber who publishes under Jordan L. Boyd-Graber in NIPS.

Written by Marek (http://www.marekrei.com/blog/author/marek/) — Posted in Uncategorized (http://www.marekrei.com/blog/category/uncategorized/)

9 comments

JUNE 30, 2016 - 5:51 PM
Jay

Do you make the data public? I am curious how Imperial College London performs for machine learning

> JUNE 30, 2016 - 6:29 PM
> Marek (http://www.marekrei.com/blog/author/marek/)
>
> The full data on organisations is quite noisy at the lower ranks at the moment, as it is extracted from pdfs and then post-processed with manual rules. It still contains a long tail of alternative spellings and entries that are not institutions at all (eg College Park). Imperial College London comes up with 7 entries in there. Although worth noting that I'm only looking at 6 specific conferences, and Imperial seems to be publishing in somewhat different areas.

JUNE 30, 2016 - 9:05 PM
Jordan Boyd-Graber (http://boydgraber.org)

I'm someone who tries to straddle both ML and NLP, but it seems that my ML persona isn't getting matched for NIPS. There I'm Jordan L. Boyd-Graber:

https://papers.nips.cc/author/jordan-l-boyd-graber-6725 (https://papers.nips.cc/author/jordan-l-boyd-graber-6725)

(This should give me some purple on the bar graphs and give me a little more ML cred.)

> JUNE 30, 2016 - 9:48 PM
> Marek (http://www.marekrei.com/blog/author/marek/)
>
> Thanks! Indeed, I'm not catching alternative names for authors at the moment. I will update it soon and add a fix for your name.

JUNE 30, 2016 - 10:39 PM
Jason Eisner (http://cs.jhu.edu/~jason)

How about including TACL? It's a journal, but deliberately set up to be another mechanism for publishing normal ACL-style papers, so leaving it out of the analysis is strange. The format is essentially the same as ACL/NAACL/EMNLP/EACL, and you get to present the work at one of those conferences. Downloading and scraping the papers should be no different than for ACL. Whether you submit via TACL or directly via the conferences is as much a matter of when the deadlines fall as anything else. (Although TACL papers arguably should count a bit more: they generally get more thorough reviews, are often required to make revisions for final acceptance, and tend to be longer.)

There's also a question of whether long-form journal papers (JMLR, CL, etc.) should be included in measures of productivity. Perhaps those are often just synthesizing and expanding previously published conference papers? – but I'm not sure.

Of course, I hope that no one optimizes for your ranking.

JUNE 30, 2016 - 10:59 PM
Marek (http://www.marekrei.com/blog/author/marek/)

The 6 conferences I chose simply based on which sources I personally follow the most. I completely agree that there are many other conferences and journals that could be included: TACL, COLING, CoNLL, *Sem, IJCAI, IJNLP, LREC, JMLR, CL, CIKM, AAAI, WWW, etc.
I intend to post an update at the end of the year, and will include a longer list of conferences. Feel free to suggest additional sources which I haven't listed yet.

JULY 1, 2016 - 3:58 AM
Wei Xu (http://www.cis.upenn.edu/~xwe/)

I second Jason. TACL is essentially equal to ACL/NAACL/EMNLP/EACL; it is quite different from COLING, CoNLL, *Sem, IJCAI, IJNLP, LREC, JMLR, CL, CIKM, AAAI, WWW, etc, and much more right in the center of NLP research. I would recommend anyone interested in NLP to follow TACL papers (if not more closely) in addition to ACL/NAACL/EMNLP/EACL.

JULY 1, 2016 - 11:39 AM
Matthias Gallé

Is this long papers only or long+short?

JULY 1, 2016 - 2:17 PM
Marek (http://www.marekrei.com/blog/author/marek/)

Looking at both long and short papers.

Theano Tutorial (http://www.marekrei.com/blog/theano-tutorial/)