IR DON'T KNOW S***** (SEARCH)

David D. Lewis
Cyxtera Technologies
Dallas, TX, USA
desires2018paper@davelewis.com

ABSTRACT

Numerous core tasks within information retrieval and natural language processing are often viewed as solved and unworthy of further research attention. I suggest we know less than we think we do, and that further research would have substantial benefits to practitioners.

1 INTRODUCTION

Many traditional information retrieval (IR) tasks are broadly viewed as "solved". Research on tokenization, phrase formation, stemming, term weighting, relevance feedback, clustering, and other basics rarely appear in major IR conferences and journals unless dressed up in new clothes. (Example of new clothes from SIGIR 2018: term weighting that's differentially private, ranking that's streaming, anything as long as it's deep neural networks,....)

Basic natural language processing (NLP) tasks of substantial interest to IR, including morphological analysis, collocation finding, language identification, lightweight discourse segmentation, and part of speech (POS) tagging face similar, though perhaps less extreme, perceptions of completeness. Manning has critically examined the notion that POS tagging is solved [1].

The illusion of completeness is quickly dispelled when one works in application areas such as electronic discovery, personal information retrieval, and enterprise search. IR software in these areas must deal with documents (and partially textual records) which vary wildly in size, vocabulary, format, genre, structure, duplication, transduction error rates, and numerous other characteristics. Multilingual datasets, and multilingual documents, are common. Not only do documents vary within datasets, but datasets vary widely from client to client of a software business. Language itself evolves.

Practitioners know that rather little of current IR and NLP technology is robust in the face of these variations (Figure 1). Attempts to achieve robustness by combining best practices across variable inputs (e.g. different human languages) occupy much more engineering effort for companies in these spaces than "interesting" machine learning.

Here are four brief examples of research that, if carried out, would likely bring substantial benefit to industrial practitioners:

- IR techniques applied to email threads currently either treat each message as a separate document, or concatenate all unique text together. The extensive body of research on XML document retrieval could helpfully be adapted to the tree structure of email.
- Combining NLP strategies for morphological induction with those for collocation finding might well lead to a universal, language-neutral but language-respecting, generator of indexing units. Engineers in charge of maintaining masses of stemmers and phrase formation rules in multilingual software would view such a development with delight.
- Modern supervised learning systems excel with moderate to large training sets. Users, however, will happily apply supervised learning capabilities to a single positive training document. This results in memorization at best, and complete failure at worst. Search systems, on the other hand, will treat a positive example as a query, and use collection-based term weighting to achieve reasonable effectiveness. Traditional relevance feedback algorithms (mostly hacks on naive Bayes) straddle the two regimes, dominating in the regime of tiny training sets. Surely all these approaches should be combined.
- Frequently updated document populations make collection-based term weights an engineering headache when used upstream of cached machine learning analyses (e.g. latent spaces). To what extent are language-specific but collection-independent weights sufficient?

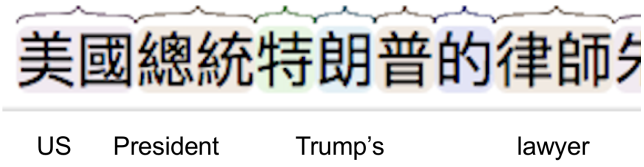

Figure 1: The most well-known name in the world, mistokenized (as three characters) and mistagged (as adverb, suffix, and common noun) by a leading commercial NLP tool.

The illusion that basic IR problems are solved results from both the narrowness of data sets used in most research studies, and from the narrow bandwidth of communication from applications back to the research community. The former problem requires creative solutions, as the diverse data of interest is expensive to assemble and frequently implicates privacy and commercial concerns. The latter problem should be helped by conferences such as this one.

REFERENCES

[1] Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In *International conference on intelligent text processing and computational linguistics*. Springer, 171–189.

DESIRES 2018, August 2018, Bertinoro, Italy