Mapping the scientific narrative

Bradley Davidson*, Robert Malouf†

*Ogilvy CommonHealth Worldwide, 440 Interpace Parkway, Parsippany, NJ 07054, brad.davidson@ogilvy.com

†Department of Linguistics and Asian/Middle Eastern Languages, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182, rmalouf@mail.sdsu.edu

You shall know a word by the company it keeps. (Firth 1957)

Introduction

The scientific literature in any (sub-)domain constitutes a kind of an ongoing narrative constructed jointly by a community of researchers using a specialized language among themselves. This goes beyond the use of technical terminology and biomedical jargon (UMLS) or English for Specific Purposes, and the narrower the subfield, the subtler the linguistic distinctions. Understanding these differences is vital for accessing the scientific narrative.

Corpus analysis

Using the tools of corpus linguistics and computational lexicography, we can analyze large quantities (on the order of hundreds of millions of words) of domain-specific text. One primary tool of corpus linguistics is the **concordance**:

```
evels are associated with <increased> disability progression i
enewed disease activity , <increasing> disability , or emergent
gression primarily due to <increasing> disability . The long-t
ent and ultimately become <increasingly> disabled . Patients freq
meters of ongoing or even <increasing> disease activity would b
meters of ongoing or even <increasing> disease activity would b
ew lesions decreases with <increasing> disease duration in adul
rophy rate decreased with <increasing> disease duration , where
nd high dietary vitamin A <increases> disease severity in the
iveness of therapy with a <increasing> disease duration or EDSS
```

This can reveal surprising patterns -- for example, in papers on multiple sclerosis, the verb *increase* occurs with undesirable direct objects like *disability* or *disease activity*, while in the monoclonal antibody literature *increase* also occurs with desirable outcomes:

```
berine showed significant <increased> effects on cell death wh
in-based therapies ) with <increased> effects of NOD2 variants
specificity and linked to <increasingly> efficacious therapies .
are systems intensifies , <increasingly> efficacious and cost-con
of up to 6 times produced <increased> efficacy without observe
ed in this study . As the <increased> efficacy of SF1126 versu
60 model , hEBV321 showed <increased> efficacy as compared to
ctively translate into an <increased> efficacy at the postsyna
h warfarin , they display <increased> efficacy with a good saf
r therapies , not only to <increase> efficacy against cancer
```

Synonym sets

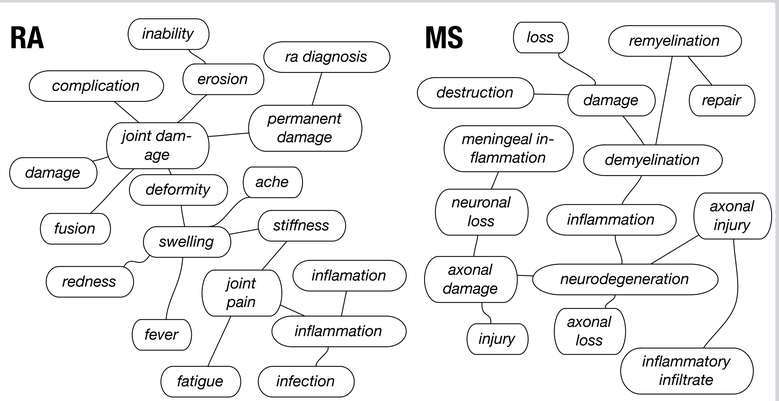
A concordance offers a summary of a word's meaning (in Firth's sense):

Activated microglia may release free radicals, nitric oxide, and proteases that may contribute to tissue _____.
The participant acknowledges that he/she has no right to lodge _____ claims against the organisers.
This inflammation results in an improved outlook for control of JCV while causing associated inflammatory _____ in the brain.
During the relapsing-remitting phase of the disease, _____ slowly accumulates over many years.

Going beyond simple word counts, information-theoretic measures of association combined with deep syntactic analysis allow automatic extraction and visualization of a domain-specific thesaurus (Lin 1998, Curran and Moens 2002). We reduce the corpus to a set of dependency triples:

(*accumulates* SUBJ *damage*), (*causing* OBJ *damage*), (*associated* MOD *damage*), (*inflammatory* MOD *damage*), ...

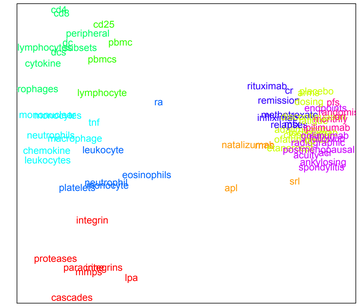
Words with similar relation profiles likely have similar meanings.



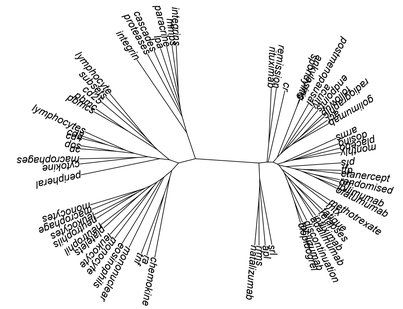
These synonym sets provide a high-level overview of the way that language is being used in a narrowly focused corpus which in turn can help the analyst find differences in word usage between that domain and biomedical literature in general.

Semantic spaces

Finally, broader semantic patterns of word meanings and language use can be found using vector space analysis and non-negative matrix factorization (Pauca et al. 2004, Turney and Pantel 2010, Utsumi 2010). This technique maps words into locations in a semantic "space":



The closeness of two words in the **semantic space** is a measure of the similarity of the larger contexts in which the two words tend to occur, and the structure of the semantic space provides a basis for comparing the development of word meanings across domains and across time.



References

Curran, J and M Moens. 2002. "Improvements in automatic thesaurus extraction." In Proceedings of the Workshop on Unsupervised Lexical Acquisition, 59-66.

Firth, JR. 1957. "A synopsis of linguistic theory 1930-1955." In FR Palmer (ed.) Selected Papers of JR Firth 1952-1959, Longman, 1968.

Lin, D. 1998. "Automatic retrieval and clustering of similar words." In COLING-ACL (1998), 768-774.

Pauca, VP, F Shahnaz, M Berry, and R Plemmons. 2004. "Text mining using non-negative matrix factorizations." In Proc. SIAM Inter. Conf. on Data Mining.

Turney, PD and P Pantel. 2010. "From frequency to meaning: Vector space models of semantics." Journal of Artificial Intelligence Research 37:141-188.

Utsumi, A. 2010. "Evaluating the performance of nonnegative matrix factorization for constructing semantic spaces: Comparison to latent semantic analysis." In Proceedings of 2010 IEEE International Conference on Systems, Man and Cybernetics, 2893-2900.