POSITION PAPER

Robert Malouf
San Diego State University

Public interactions that a generation ago were ephemeral are now mediated by the internet and leave a permanent and accessible record. Combined with large-scale natural language processing and text analysis, this has created unprecedented opportunities for scientists studying human linguistic behavior. Until recently, much of the text available on the WWW was either professionally edited, or followed the conventions of edited text. Newspapers, magazines, corporate and government publications, academic papers, and even personal and hobby sites produced by amateurs follow fairly rigid standards for formatting, style, and orthography. More importantly, they are intended to be read by a large anonymous audience which shares only the most general public context. This makes them easy to process with relatively little specific background knowledge, both for human readers who may be referred to the site by a search engine, and for automated methods such as question answering or text mining systems.

On the other hand, in online environments such as discussion forums, social networking sites, and chat rooms, where content is created by the users themselves, the use of language is very different. Unlike edited text, informal web texts are typically conversational, are often non-standard or idiosyncratic, and are highly contextualized, depending on rich background of shared knowledge and assumptions. The only audience the text is intended for is the immediate participants in the online discussion at the time the text was produced. Informal web texts pose new and interesting problems for text processing techniques which have been developed for more traditional edited text genres.

In a related development, researchers have access to enormous quantities of published literature in all scientific disciplines, most especially in the biomedical arena. 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 (as documented in resource like 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.

One thing that both scientific sub-fields and online communities have in common is that their use of language is targeted at an audience of 'insiders' and must be approached accordingly. Using the tools of corpus linguistics and computational lexicography, we can analyze large quantities (on the order of hundreds of millions or billions of words) of domain-specific text. One primary tool of corpus linguistics is the concordance, an index of every occurrence of a word or phrase in context. 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, such as *efficacy*.

A concordance offers a summary of a word's meaning: "You shall know a word by the company it keeps." (Firth 1957). 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', pairs of words linked by a grammatical relation. If we assume that words with similar relation profiles likely have similar meanings, we can use the network

of grammatical relations to induce a semantic network of related terms and concepts. 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.

Finally, broader semantic patterns of word meanings and language use can be found using techniques such as 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. As an example, we can map project neighborhoods in a city into a semantic space, reflecting similarities and difference in the way residents talk about them. This allows us to produce conceptual city maps in which distances are derived from neither physical proximity nor demographic similarity, but rather from the subjective role that neighborhoods play in the popular imagination as reflected in the text. This map of a city's cultural landscape can be useful to both residents and to researchers. For example, areas which are likely targets of gentrification may look different in the cultural space than other areas which objectively have very similar attributes. Or, by overlaying cultural maps of different cities, we can match up equivalent regions, allowing someone to find a neighborhood in an unfamiliar city which plays a similar role to a neighborhood in a more familiar one.