LA-UR- 04-3190

Approved for public release;
distribution is unlimited.

Title: TOWARDS A SEMANTIC LEXICON FOR BIOLOGICAL LANGUAGE PROCESSING

Author(s): Karin Verspoor, [redacted], CCS-3

Submitted to: ISMB BioLINK, Glasgow, Scotland, July 29, 2004

Los Alamos
NATIONAL LABORATORY

Towards a Semantic Lexicon for Biological Language Processing

Karin Verspoor
Los Alamos National Laboratory
verspoor@lanl.gov

It is well understood that natural language processing (NLP) applications require sophisticated lexical resources to support their processing goals. In the biomedical domain, we are privileged to have access to extensive terminological resources in the form of controlled vocabularies and ontologies, which have been integrated into the framework of the National Library of Medicine's Unified Medical Language System's (UMLS) Metathesaurus. However, the existence of such terminological resources does not guarantee their utility for NLP. In particular, we have two core requirements for lexical resources for NLP in addition to the basic enumeration of important domain terms: representation of *morphosyntactic* information about those terms, specifically part of speech information and inflectional patterns to support parsing and lemma assignment, and representation of *semantic* information indicating general categorical information about terms, and significant relations between terms to support text understanding and inference (Hahn et al, 1999). Biomedical vocabularies by and large commonly leave out morphosyntactic information, and where they address semantic considerations, they often do so in an unprincipled manner, for instance by indicating a relation between two concepts without indicating the type of that relation.

But all is not lost. The UMLS knowledge sources include two additional resources which are relevant – the SPECIALIST lexicon, a lexicon addressing our morphosyntactic requirements, and the Semantic Network, a representation of core conceptual categories in the biomedical domain. The coverage of these two knowledge sources with respect to the full coverage of the Metathesaurus is, however, not entirely clear. Furthermore, when our goals are specifically to process biological text – and often more specifically, text in the molecular biology domain – it is difficult to say whether the coverage of these resources is meaningful. The utility of the UMLS knowledge sources for medical language processing (MLP) has been explored (Johnson, 1999; Friedman et al 2001); the time has now come to repeat these experiments with respect to biological language processing (BLP). To that end, this paper presents an analysis of the UMLS resources, specifically with an eye towards constructing lexical resources suitable for BLP. We follow the paradigm presented in Johnson (1999) for medical language, exploring overlap between the UMLS Metathesaurus and SPECIALIST lexicon to construct a morphosyntactic and semantically-specified lexicon, and then further explore the overlap with a relevant domain corpus for molecular biology.

The UMLS as a Lexical Knowledge Source

There have been several investigations of the UMLS as a lexical knowledge source. McCray et al (2001) evaluate the nature of strings in the UMLS Metathesaurus with respect to their likelihood of appearing in a natural language corpus. They found that only 10% of the strings in the Metathesaurus occurred in their MEDLINE corpus (representing one year of MEDLINE abstracts), but were able to identify some properties associated with the strings that could be used to filter out strings that are unlikely to occur naturally in a corpus. While the authors suggest that occurrence of a term in the

Metathesaurus opens the possibility of accessing more extensive domain knowledge about that term, they do not explore the nature of that domain knowledge for the terms they find in their corpus, and do not explore the overlap of those terms with other UMLS resources.

Friedman et al (2001) quantitatively compare a lexicon developed manually for their MEDLEE system with a lexicon derived automatically from the UMLS, with respect to the task of processing clinical information in patient reports. They found the UMLS-derived lexicon to lead to poor performance relative to their own lexicon. The results do not, however, invalidate the UMLS as an important source of lexical information, as they may be a reflection of the completeness of the existing MEDLEE lexicon for the task evaluated. The authors argue that using the UMLS can substantially reduce the manual effort in constructing a lexicon.

Johnson (1999) explores the construction of a lexical resource from the UMLS in support of processing of medical narrative, specifically utilizing a corpus of discharge summaries from hospital visits. Johnson explores the overlap between the Metathesaurus, the SPECIALIST lexicon, and a domain corpus, and presents some strategies for handling semantic ambiguities that arise during the mapping of terms in the different UMLS resources. Johnson found that while 79% of the distinct lexical forms in his corpus occurred in the SPECIALIST lexicon, only 38% of those forms occurred in the semantic lexicon of more than 75,000 entries derived from intersecting the Metathesaurus and the SPECIALIST lexicon – so, only 38% of terms in the corpus could be expected to have both morphosyntactic and semantic information derived from the UMLS. Johnson points out this may reflect the fact that the Metathesaurus may contain many complex medical terms that should not be considered lexical items, and that furthermore may successfully be incorporated into the lexicon by assuming that they are nouns.

Methods

We follow Johnson (1999) and explore the overlap in the UMLS Metathesaurus and the SPECIALIST lexicon to establish a baseline semantic lexicon, and then investigate its relevance for a corpus in the molecular biology domain. We utilize the 2003AC UMLS release. As our domain corpus, we utilize 29,514 full text articles from the Journal of Biological Chemistry (JBC), spanning the years 1998-2002, originally obtained for the 2003 BioCreAtIvE competition. While we realize that this is not a sample representative of the full domain of molecular biology, it is representative of a significant portion of that domain, and the results on JBC texts should be indicative of the coverage of our semantic lexicon for this domain. We felt it preferable to use a corpus of full text articles rather than a corpus of abstracts derived from MEDLINE in order to more completely assess coverage of the relevant language.

The steps for building and evaluating our semantic lexicon are as follows:

- Lexemes in the SPECIALIST lexicon are matched to terms in the Metathesaurus. We load in all the strings represented in the SPECIALIST LRAGR file, and attempt to match Metathesaurus strings extracted from the MRCON file to these strings. This is done by considering different kinds of matches:
    - Exact match
    - Match after uppercasing the first letter of the SPECIALIST string

- Match after uppercasing the first letter of each word of the SPECIALIST string
- Match after uppercasing the entire SPECIALIST string
- Other case insensitive match
- Match (any of the above types) after stripping the Metathesaurus string of ", NOS" or "<1>", "<2>", etc. at the end
- Finally, consider whether each of the constituent words of a Metathesaurus string occurs in the SPECIALIST lexicon (after removal of words consisting of all numbers or punctuation), in order to assume a compositional analysis of the term

- Filter the resulting lexicon (a subset of the original SPECIALIST lexicon tied to specific Metathesaurus terms) by removing any terms for which the corresponding Metathesaurus string is not associated with a semantic type through one of its associated concepts. There are concepts for which the UMLS does not provide semantic information, and therefore they do not satisfy our lexical constraints requiring both morphosyntactic and semantic information.
- Search the domain corpus for occurrences of any lexical variant of each term in our semantic lexicon (obtaining lexical variants from the UMLS lexical tools), and track any matches. We also consider the overlap of the most frequent terms in the corpus with the lexicon.

Results

As of submission of this abstract, we have only completed the first step of the above methods. The remaining steps will be completed prior to submission of the final version of the abstract for the workshop and therefore incorporated into that version of the abstract. Thus far, as shown in the table below, our results indicate that the proportion of Metathesaurus terms directly occurring (through some matching paradigm) in the SPECIALIST lexicon is in fact slightly less than Johnson's (1999) finding of 12% at 8.2%. This is due to the incredible growth in the Metathesaurus in the past few years; Johnson reports finding 630,658 unique strings in the Metathesaurus, while we found 1,959,516 unique strings. The SPECIALIST lexicon has grown as well (from 164,850 distinct lexical forms to 292,979), but clearly not at pace with the Metathesaurus. This result is in line with Johnson's observation that many of the terms in the Metathesaurus are probably not appropriate for recording directly in the SPECIALIST lexicon. However, upon inspection of the constituent structure of Metathesaurus terms, we found that for a large proportion of terms (79%) each of the constituent members of the terms could be found in the SPECIALIST lexicon. This opens the possibility of a compositional analysis for many Metathesaurus terms, though it doesn't address the assignment of semantic type to the term as a whole.

Exact matches	58,918	3.0%
First letter uppercase	67,765	3.5%
First letter, all words uppercase	13,922	0.7%
Entire string uppercase	12,961	0.7%
Other case insensitive match	1,982	0.1%
Stripped term matches	5,945	0.3%
Total direct matches	**161,493**	**8.2%**
Constituent matches	1,548,389	79.0%
Total matches	**1,709,882**	**87.3%**

We will establish the semantic nature of our lexicon, and the overlap with the domain corpus through completion of our analysis. There analysis will be completed in the next few weeks.

Conclusions

As this work is as yet incomplete, we cannot present definitive conclusions. However, we expect to find sufficient overlap with our derived semantic lexicon to justify the use of the UMLS resources as a starting point for a lexicon for Biological Language Processing.

There remain questions about the utility of the UMLS Semantic Network for BLP. Although we have established a core lexicon for which we have the basic required lexical information – morphosyntactic and semantic information – we have not investigated any potential shortcomings of the UMLS Semantic Network. There are 135 semantic types and 54 relationship types represented in the 2003AC version of the Semantic Network; the number of types is quite small given the complexity of the biomedical domain, and this begs the question of whether it adequately characterizes the semantic distinctions needed for BLP. In contrast, the Gene Ontology resource (Ashburner et al, 2000) contains over 16,000 concepts grouped hierarchically and therefore in principle represents a much more fine-grained semantic breakdown of the domain. The GENIA ontology under development (Ohta et al, 2002) is focused on cell signaling reactions in humans and as such characterizes concepts specific to those processes, again likely to be much more fine-grained than the broad UMLS ontology. The relative utility of different ontologies should be investigated.

References

Ashburner, M; Ball, C.A.; and Blake, J.A. et al [2000]. "Gene Ontology: Tool for the Unification of Biology", Nature Genetics, v. 25:1, pp 25-29.

Friedman, Carol, Hongfang Liu, Lyuda Shagina, Stephen Johnson, George Hripcsak [2001]. Evaluating the UMLS as a Source of Lexical Knowledge for Medical Language Processing. Proc. AMIA 2001; 189-193.

Hahn, Udo, Martin Romacker, Stefan Schulz [1999]. How Knowledge Drives Understanding – Matching Medical Ontologies with the Needs of Medical Language Processing. Artificial Intelligence in Medicine, 15:25-51.

Johnson, Stephen B [1999]. A Semantic Lexicon for Medical Language Processing. Journal of the American Medical Informatics Association, 6:3, 205-218.

McCray, Alexa T., Olivier Bodenreider, James D. Malley, Allen C. Browne [2001]. Evaluating UMLS Strings for Natural Language Processing. In the Proceedings of the AMIA Annual Symposium 2001; 448-452.

Ohta, Tomoko, Yuka Tateisi, Jin-Dong Kim [2002]. GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In the Proceedings of the Human Language Technology Conference (HLT 2002).