Adding a Medical Lexicon to an English Parser

Peter Szolovits, PhD
MIT Laboratory for Computer Science, Cambridge, MA

ABSTRACT

We present a heuristic method to map lexical (syntactic) information from one lexicon to another, and apply the technique to augment the lexicon of the Link Grammar Parser with an enormous medical vocabulary drawn from the Specialist lexicon developed by the National Library of Medicine. This paper presents and justifies the mapping method and addresses technical problems that have to be overcome. It illustrates the utility of the method with respect to a large corpus of emergency department notes.

INTRODUCTION

Computer-readable clinical data are today largely stored as unstructured text.[1] Although a few projects report impressive progress toward turning such text into structured data [2,5], much remains to be done to capture a comprehensive representation of the content of medical texts by computer.

Because many other research and application communities face the same problem, we have an opportunity to share tools with others to help parse, formalize, extract and organize meaning from text. A practical impediment to such sharing, however, is that most language processing tools built by those outside medical informatics have no knowledge of the medical lexicon, and therefore treat a large fraction of medical words as if they bore no syntactic information. This deficiency makes such tools practically inapplicable to medical texts. For example, in trying to parse unstructured text notes recorded in the Emergency Department of a large urban tertiary care pediatric hospital (ED), the Link Grammar Parser (LP) [3,6] was able to recognize only 5,168 of the 13,689 distinct words[2] (38%) in a half-million (494,762) word corpus. This is despite the fact that LP has one of the larger lexicons distributed with such tools, holding syntactic definitions of 49,121 words (including about 1,000 short phrases). Except as noted otherwise, we treat phrases exactly the same as words.

One of the richest available sources of medical lexical information today is the UMLS's Specialist Lexicon [4]. The 2001 version, which we use here, contains lexical information on 235,197 words (including 75,121 short phrases). It is a more complete lexicon than others for purposes of analyzing medical text. For the ED corpus, for example, 7,328 of the distinct words (54%) appear directly in the Specialist Lexicon. Of these words, 2,284 do not appear in the LP lexicon, yet account for 17% of the distinct words in the corpus.[3]

Our goal in this paper is to describe a method for expanding the usable lexicon of LP (the *target*) by adding lexical definitions of words present in the Specialist Lexicon (*source*) but not in LP. Because the nature of lexical descriptions varies greatly among different lexicons, one cannot simply copy definitions from one to another. Methods similar to those presented here should be applicable to other pairs of language processing systems that offer significantly different coverage of vocabularies. The next section describes lexical information available for categorizing words in both the Link Parser and Specialist lexicons, and the method by which we map Specialist words to Link Parser words. Later, we describe the results of this mapping and show a dramatic increase in the lexicon of the LP. Finally, we discuss ways to address inconsistency between lexicons, capitalization, still missing words, and the treatment of words with unique lexical descriptions.

METHODS

Every lexicon uses some language of lexical descriptors to specify lexical information about each word or word sense. Although a language processing system's knowledge about a particular word may be quite extensive, often the bulk of the knowledge is about the meaning rather than the lexical constraints

[1] Advanced clinical information systems store at least some of the following types of data in coded and structured form: laboratory data, recorded signals and images, pharmacy orders, and billing-related codes. If other clinical data are captured at all, items such as medical history, physical examination findings, progress notes, discharge summaries, radiology and pathology notes, etc., are nevertheless stored simply as text.

[2] This assumes that different capitalizations of a word are nevertheless instances of the same word. If we distinguish different capitalizations of a word, there are 17,372 distinct words in the corpus. We discuss issues of capitalization below.

[3] It may be surprising that even a lexicon developed for medical use is missing 46% of the distinct words in this corpus. We examine this in the Discussion section.

on the word. Therefore, lexical descriptions tend to be relatively small, and many words, though they may differ greatly in meaning, will have the same lexical description. Central to our method of mapping lexical structure is the notion of *indiscernibility*. Two words are indiscernible in a specific lexicon just in case that lexicon assigns the same lexical descriptors to the words.

Heuristic mapping. The basic idea of mapping is simple: Assume that *w* is a word of the source lexicon for which no lexical information is known in the target lexicon. If there is a word *x* in the source lexicon that is indiscernible from *w* and if *x* has a lexical definition in the target lexicon, then it is at least plausible to assign the same lexical definition in the target lexicon to *w* as *x* has.

This simple picture is complicated by a handful of difficulties: lexical ambiguity and inconsistency between different lexicons, different lexical treatment of capitalized and lower-case words, words whose lexical descriptions are unique, and words that don't appear at all in either lexicon. To make mapping practical, we have had to overcome, at least in part, each of these. The following sections describe in detail the lexical descriptors available for words in both the Specialist and LP lexicons, and then address each of the above-listed difficulties, in turn.

Structure of the Specialist lexicon. The following lexical knowledge relevant to our approach is encoded in Specialist. We extract the information from the listed relational tables.

1. *Part of speech*. (`lragr`)
2. *Agreement/Inflection Code*. First, second and third person; singular and plural; tense and negation (for verbs, modals and auxiliaries); count/uncount for nouns and det's. (`lragr`)
3. *Complements*. A complex system for describing the types of complements taken by verbs, nouns and adjectives, including linguistic types of the various complementation patterns, prepositions, etc. (`lrcmp`)
4. *Position and modification* types for adjectives and adverbs. (`lrmod`)
5. *Other features.* (`lrmod` for adjectives and adverbs)

We ignore lexical markers that relate only to "closed" classes of words such as pronouns, because we expect that such words will already be defined and lexically described in the target lexicon. We also ignore inflectional markers because both source and target lexicons already include inflected word forms. For example, both "run" and "running" appear as distinct entries in both lexicons.

Both Specialist and LP usually contain separate entries in the lexicon for a word that has senses corresponding to different parts of speech. For example, "running" is marked in Specialist as either a verb present participle or as a third-person count or uncount noun. LP's analysis makes it either a verb or a gerund. Consequently, we will treat each such word/part-of-speech pair as a *word sense*, and apply our heuristic mapping method to such word senses. Note that neither lexicon contains distinct entries for traditional word senses such as the distinction between a bank (financial institution) and a (river) bank, perhaps because these are distinctions in meaning, not lexical structure. Specialist has a total of 246,014 word senses, and LP has 58,926.

Within the Specialist lexicon, we consider two word senses to be indiscernible if and only if they have exactly the same set of lexical descriptions according to the knowledge sources enumerated here. For example, though the words "execute" and "cholecystectomize" may share little in meaning, they do have exactly the same lexical markings in Specialist. Both are verbs, marked as "infinitive" and "pres(fst_sing, fst_plur, thr_plur, second)", and admit complement structures "tran=np" or "ditran=np, pphr(for, np)".

The Specialist lexicon contains 3,357 distinct sets of indiscernible entries among its 246,014 total word senses. Fully 2,574 of these sets are singletons, meaning only one entry is contained in the set, because the lexical markings of that entry are unique. (See discussion, below.) By contrast, the largest indiscernible set contains 64,149 entries—about ¼ of the entire lexicon. Figure 1 shows the distribution of bin sizes after excluding singletons. Thus, there are 417 indiscernible sets of sizes 2 or 3, 200 of size 4-10, 85 of size 11-32 … and 2 of size > 32,000.

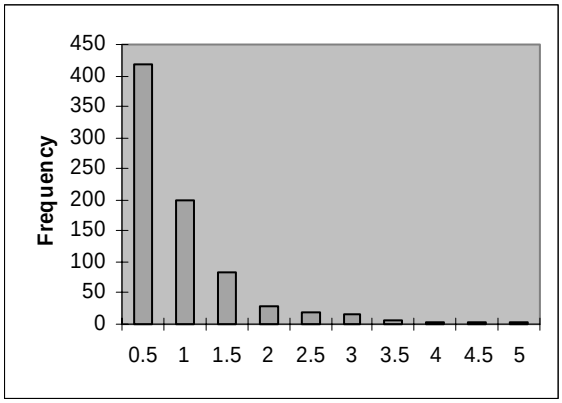

Fig. 1. Histogram of the number of indiscernible lexical sets in Specialist plotted against the size of the set. Abscissa is $\log_{10}$ of the maximum set size to fall into that bin.

Successive bins correspond to units of 0.5 in $\log_{10}$ of the bin size.

Structure of the LP lexicon. The LP lexicon associates with each word sense a complex Boolean formula of features. These are drawn from a set of 104 basic features, augmented by additional markers that specify whether a matching feature is expected on a word to the left or right in the sentence, whether a match on this feature is optional, required or possible but penalized, and many additional sub-specializations of the basic features to enforce further lexical constraints. Indiscernibility of two lexical descriptions in LP should properly be based on computing equivalence classes among these formulae. This computation is rather complex, however, because fragments of formulae that conjoin constraints pointing in the same direction (i.e., to the left or right) do not commute. Fortunately, almost all sets of words that share the same formula are defined together in the LP lexicon; therefore, we can assume that two words are lexically indiscernible just in case they are defined by the exact same formula.

The LP lexicon contains 1,324 sets of indiscernible word senses. 767 of these are singletons, and the largest set contains 11,107 word senses.

Mapping Formalism. Let W be the set of word senses (word × part-of-speech) in the source lexicon and V be the set of word senses in the target lexicon. For each $w \in W$, let $X_w = \{x \mid x$ is indiscernible from w in the source lexicon$\}$. Define $f(v)$ to be the lexical formula for v in the target lexicon, if v is defined there, or $\perp$. Further let $D_w = \{f(x) \mid x \in X_w$ and $f(x) \neq \perp\}$; this is the set of definitions in the target lexicon that belong to word senses indiscernible from w in the source lexicon. Clearly, we will want to associate one of the definitions in D_w with w via our heuristic mapping, but if there are several, which one? Define $I(d) = \{v \mid f(v) = d\}$; this is the set of indiscernible word senses in the target lexicon that share the lexical description d. For each $d \in D_w$, we compute the number of word senses in common[4] between $I(d)$ and X_w, and choose the definition that yields the largest overlap: $m(w) = \text{argmax}_{d \in Dw} \|X_w \cap I(d)\|$ is then the lexical descriptor of w in the target lexicon. Ties go to the largest $\|I(d)\|$. Figure 2 shows a schematic version of the mapping algorithm.

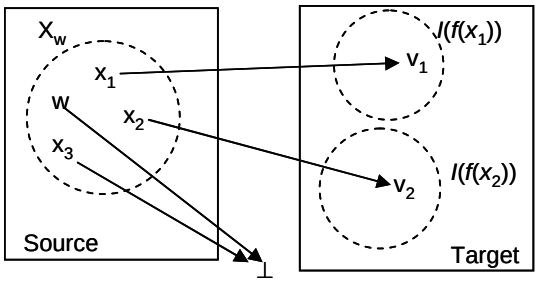

Figure 2. Our heuristic maps w to the definition of a word sense in V, say, $d(v_2)$. Each x indiscernible from w potentially identifies a set of indiscernible word senses in the target lexicon, $I(f(x))$. We choose the one that has the largest number of word senses in common with X_w.

For example, consider the word "cholecystectomize", a verb defined as described above in Specialist, but unknown to LP. In Specialist, it is indiscernible from 20 other words, including "adduce", "admire", "appropriate", "execute", … The verb senses of these words, when looked up in LP as verbs, yield four distinct formulae for D_w. For each, we compute the set of LP-indiscernible words whose verb senses go with that formula. Only one of these word sets (the largest) has an intersection greater than 1 with the words grouped by Specialist with "cholecystectomize", hence the LP formula for that set is the appropriate definition of "cholecystectomize" in LP.

RESULTS

The mapping process is quite productive. From the 246,014 words senses in the Specialist lexicon, the mapping process constructed 200,264 new entries (word senses) for the LP lexicon. This approximately *quintuples* the size of that lexicon, and certainly adds a huge number of lexical definitions for medical terms and phrases. 73,780 of the new word senses are actually phrases (containing a space, hyphen or comma), and many of these phrases bear no specific lexical information in Specialist that is not obvious from their component words. For example, "chronic relapsing pancreatitides" is simply marked as a third person plural count noun, which is also true of "pancreatitides". For now, we have chosen to retain these phrases in the augmented lexicon, but may choose to drop them in the future.

How good is the mapped definition of the remaining 126,484 new word senses introduced into the target lexicon? To answer this question by an exhaustive study of each new definition would be nearly as difficult as creating that many new definitions for LP from scratch. Nevertheless, we can argue both from the logic of the mapping process and from an

[4] One of these sets is in the source lexicon, whereas the other is in the target. Two word senses are "in common" when they are both senses of the same word and when the part of speech descriptions from both lexicons are consistent with each other. For example, LP contains eight markers that are all consistent with Specialist's "noun."

examination of its results that, at least in the large, the mapping seems to make reasonable definitions.

All of the new word sense definitions fall into only 59 indiscernible sets in LP. The five largest of these sets account for 181,653 of the 200,264 new word sense definitions (or 109,406 of the 126,484 new definitions if we exclude phrases). These include (1) third person singular uncount nouns ("dissuasion, genaconazole, vantocil"), (2) third person plural count nouns ("pinealomas, tracheotomies, perfects"; the latter seems to be an error, but is so defined in Specialist), (3) positive stative adjectives that are either predicative or one of the forms of attributive known to Specialist ("vasomotory, atherogenic, seminiferous"), (4) third person singular count or uncount nouns ("antilipopolysaccharide, chrysiasis, choline"), and (5) third person singular count nouns ("diplococcus, lumbricus, milliunit"). Five other sets define between one and three thousand related words, e.g., past tense or past participle verbs ("nebulized, imprecated, autosensitized"). Nearly thirty sets define fewer than ten word senses, some as few as one. For a few of these, the mapping heuristic seems to choose poor definitions, which have been manually fixed.

The largest twenty mapped sets, which account for nearly all the newly-defined word senses in LP, belong in fact to the most populous grammatical categories known to either Specialist or LP. They contain just the wealth of medical terms that we had hoped to introduce to LP's lexicon. The fact that all these words fall into only a handful of indiscernible sets suggests that many subtle syntactic distinctions for these new words may be missing (and may, indeed, have been missing even in the Specialist lexicon). Nevertheless, the most critical information about part of speech, gender and number agreement, and perhaps a few other syntactic features have been successfully transferred.

FURTHER DISCUSSION

Lexical Inconsistency. We have noted instances where LP makes finer distinctions among lexical word senses than Specialist does. Another source of potential inconsistency between these two lexicons is the following: In LP, if a word has two lexical senses, writers of the LP lexicon have a choice of whether to include the word twice in the lexicon, with separate formulae defining the two senses, or to include it only once, with a formula that is the disjunction of the formulae for the two senses. When the latter is done, it is usually impossible to tell the intended part of speech of the word, which causes some difficulties for our mapping algorithm.

Capitalization. Surprisingly many words (1,482) in the Specialist lexicon occur with multiple possible capitalizations. For about half these words (611), the lexical features of different capitalizations are actually the same. For example, both "aborigine" and "Aborigine" are marked with exactly the same lexical information. Any language processing program must handle capitalized versions of normally lower-cased words because of the convention in many Western languages to capitalize the first word of each sentence.

In the rest of the cases, different lexical information goes with different capitalizations. "DIP" is said to be a positive adjective, whereas "dip" can be a noun in third person singular or plural, count or uncount; or a verb that is either infinitive or present tense, first or second person singular or plural or third person plural. These words, spelled the same, appear to have no connection to each other. By contrast, "East" differs from "east" only in that the noun meaning of the word is marked as proper for the capitalized case. "Digitalis" is either a third person singular uncount noun or a third person plural uncount noun, whereas "digitalis" is said to be only the first of these, perhaps in error. In mapping words from one lexicon to another, we currently treat capitalized, upper and lower case versions of the same word as unrelated. Obviously, this could be improved at least for common cases.

Unique Lexical Descriptions. In discussing the Specialist lexicon, we observed that 2,574 of the 3,357 sets of indiscernible words were singletons. Although this represents only a small fraction (<1.5%) of the total number of entries in the Specialist lexicon, our mapping technique cannot work for these words. Fortunately, many of these are common English words that earn their unique lexical descriptor because of the many idiosyncratic ways in which they may be used. The verb sense of the word "take", for example, may be used in forms such as "take ill", "take stock", "take umbrage", "take to *something*", "take into consideration", "take back", "take after", and many other forms. It is not surprising that no other word will have exactly the same complement structure. "Take" and 2,106 other words whose indiscernible sets are singletons are already defined in the base vocabulary of LP. Of the remaining 468 words, 211 are actually short phrases containing words separated by spaces or hyphens, such as "cross-react" or "masking tape". These might be properly handled, at least at the lexical level, simply as a result of the lexical features of their component words.

For the remaining 257 words in this category, it seems mostly that the Specialist lexicon has over-specialized their lexical descriptions. For example, "periarthritis" is marked to take the complement "of shoulder", which is certainly a plausible complement but by no means necessary or even common. These words could be mapped successfully by eliminating their complementation features altogether, or at least by reducing their number until the word falls into a non-singleton discernibility set. We have not yet implemented this likely improvement.

Unrecognized Words. We noted that a large fraction of words in the ED corpus do not appear in the Specialist lexicon. Examination of those missing words shows that the vast majority belong to the following categories:

1. *Misspellings:* e.g., rhinnhorea, rhinnorhea, rhinorhea, rhiorrhea, and rhnorrhea all appear as incorrect variants of rhinorrhea.
2. *Proper names.* These are mostly people's names, but also include some place and institution names and medical brand names. Some brand names appear to be in the lexicon, though most are not. Some, though not most, names of places and people also occur in the lexicon, especially if they are associated with specific diseases.
3. *Numbers concatenated to units of measure or other words:* e.g., 10mmHg, 15kg, 145pm, 156HR, 15attacks, 24gauge. 2,016 of the 13,689 distinct words in the ED corpus fall into this category. That is almost 15%, or about 1/3 of the ED words not found in the Specialist Lexicon.
4. *Abbreviations not known to the lexicon:* e.g., tib, tox, trach, ul, vag, ven, yo, x10, x10d, and x38d.
5. *Prepended compounds:* Compound words consisting of a prefix such as "hyper", "non", "para", "un", etc., combined with known words. Lexical decomposition algorithms should find these, though they are not statically represented in Specialist.

Only a small fraction of these missing words should, arguably, be in the lexicon. A few examples are: "soupy", "snowbank", "popsicle", "Pedialyte".

Using a spelling corrector, a lexical analyzer that identifies and suggests lexical descriptors for the prepended compounds, and a program to recognize and expand many more abbreviations could yield a practical solution to dealing with most of these unknown words. These challenges will arise when using any lexicon, and are really orthogonal to the focus of this paper.

CONCLUSION

We have demonstrated a heuristic technique for mapping syntactic information about words defined in one lexicon to a different lexicon. We have applied this method to add syntactic descriptions of two hundred thousand medical words and phrases to the Link Parser lexicon, thereby vastly increasing its knowledge of medical terminology. We have also shown that the combined new lexicon contains nearly all of the words appearing in a half-million word corpus of notes from an emergency department, with the exceptions of the categories described above.

The translated entries from the Specialist Lexicon to LP are available for interested users to download at http://www.medg.lcs.mit.edu/projects/text/, along with some additional technical details of the mapping process that cannot fit in a short paper. We are currently studying the use of the augmented LP grammar to extract meaning from the ED corpus. An earlier version has found application in a research project that uses conversational interaction with internet users to learn new facts about a domain [1].

The approach presented here should be applicable to transferring syntactic knowledge among other lexicons as well, and could, therefore, ease the burden of equipping general-purpose language processing systems with specialized vocabularies tailored to specific domains of discourse.

Acknowledgement. This research has been supported (in part) by contract 290-00-0020 from the Agency for Healthcare Research and Quality.

REFERENCES

1. Chklovski, T. Using analogy to acquire commonsense knowledge from human contributors. AITR-2003-002. PhD Thesis, MIT AI Lab.

2. Friedman, C. A broad-coverage natural language processing system. *Proc. AMIA 2000;* 270-4.

3. Grinberg D, Lafferty J, Sleator D, *A robust parsing algorithm for link grammars*, report CMU-CS-95-125, and *Proc. Fourth Int. Workshop on Parsing Technologies*, Prague, September, 1995.

4. McCray AT, Aronson AR, Browne AC, *et al.* UMLS knowledge for biomedical language processing. *Bull Med Libr Assoc* 1993; 81(2):184-94.

5. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc.* 1994 1:142-60.

6. Sleator D, Temperley D, *Parsing English with a Link Grammar, Third International Workshop on Parsing Technologies*, August 1993.