Automated Non-Alphanumeric Symbol Resolution in Clinical Texts

SungRim Moon, MS[1], Serguei Pakhomov, PhD[1, 2],
James Ryan[3], Genevieve B. Melton, MD, MA[1,4]
[1]Institute for Health Informatics; [2]College of Pharmacy;
[3]College of Liberal Arts; [4]Department of Surgery
University of Minnesota, Minneapolis, MN

Abstract

Although clinical texts contain many symbols, relatively little attention has been given to symbol resolution by medical natural language processing (NLP) researchers. Interpreting the meaning of symbols may be viewed as a special case of Word Sense Disambiguation (WSD). One thousand instances of four common non-alphanumeric symbols ('+', '–', '/', and '#') were randomly extracted from a clinical document repository and annotated by experts. The symbols and their surrounding context, in addition to bag-of-Words (BoW), and heuristic rules were evaluated as features for the following classifiers: Naïve Bayes, Support Vector Machine, and Decision Tree, using 10-fold cross-validation. Accuracies for '+', '–', '/', and '#' were 80.11%, 80.22%, 90.44%, and 95.00% respectively, with Naïve Bayes. While symbol context contributed the most, BoW was also helpful for disambiguation of some symbols. Symbol disambiguation with supervised techniques can be implemented with reasonable accuracy as a module for medical NLP systems.

Introduction

Clinicians frequently use a wide range of shorthand expressions to maximize efficient communication in not only expressing linguistic meanings but also in representing medical information(1). In addition to large numbers of abbreviations and acronyms, a number of symbols are utilized as condensed meaning-bearing units in free-text clinical notes. Like words, acronyms, and abbreviations, these symbols, which consist mostly of non-alphanumeric characters, often have ambiguous senses. Symbol disambiguation may be considered an analogous problem to automatic word sense disambiguation (WSD). Since the antecedent or pre-processing Natural Language Processing (NLP) module can potentially deteriorate the quality of downstream processing functions of automatic NLP systems(2-4), proper resolution of symbols is necessary to ascertain the meaning of symbols and preempt errors in automated medical NLP systems.

Neither the medical NLP nor computational linguistics literature has focused upon symbol resolution to any large extent. In the biomedical domain, researchers have investigated disambiguation of gene symbols from biomedical text. In one such study, gene symbol disambiguation was performed with the goal of identifying biomedical entities(5). Computational linguists, in contrast, have been mainly interested in the meaning of words themselves and have largely ignored non-alphanumeric symbols outside of dealing with the task of sentence splitting.

In one analogous study that focused on symbol resolution in Chinese text, Hwang et al. examined resolution of three non-alphanumeric symbols ('/', ':', and '–') in the Academic Sinica Balance Corpus (ASBC), which consists of Mandarin and English symbols(6). They found seven senses for symbol '/', five senses for ':', and seven senses for '–'. They set up a rule-based multi-layer decision classifier (MLDC) utilizing applied linguistic knowledge with a statistical voting schema and used words surrounding the target words (bag-of-words, BoW) with statistical probabilities as features. This two-layer model was expanded into a three-layer model using preference scoring based on the location of characters/words(7). While this approach may be effective in some cases, rule-based classification with linguistic knowledge can serve as a bottleneck in maintaining automatic resolution systems because language is always changing and these rules must be maintained depending on characteristics of the corpus. Even if the MLDC used by these authors focused upon symbol disambiguation, this is at best an analogous application to English clinical note disambiguation. These results may not be directly transferrable to clinical notes because of the structural difference between English and Mandarin, and because of contextual difference between general documents and clinical notes. For example, English word tokens are separated by whitespace, but Mandarin word tokens are not.

For this pilot study, we selected four symbols ('+', '–', '/', and '#') and conducted a set of experiments for automated symbol sense disambiguation using clinical notes. We investigated symbol senses using the literature and annotations of a moderate-sized corpus, and then performed automated symbol disambiguation using three

supervised machine-learning classification algorithms: Naïve Bayes, Support Vector Machine, and Decision Tree classifiers).

Method

Symbol sense inventory

An initial sense inventory for the target symbols ('+', '–', '/', and '#') was created from several reference resources. From the field of computational linguistics, we utilized two textbooks: *Speech and Language Processing* and *Foundations of Statistical Natural Language Processing*(8, 9). We also identified several medical references with symbol senses including a medical dictionary (*Stedman's Medical Abbreviations, Acronyms & Symbols*(10)), medical terminological reference (*Medical Terminology* and references of approved symbols(11-13)), and references from the clinical literature (*Abbreviations and acronyms in healthcare*(14)).

The symbol sense inventory was then refined to remove unclear senses and add missing senses identified by a clinician (GM), and two linguists (JR and SP). "Literature sense" represents this initial sense inventory for the target symbols.

Experimental samples and document corpus

The document corpus for this study consisted of electronic clinical notes from University of Minnesota-affiliated Fairview Health Services (consisting of four metropolitan hospitals in the Twin Cities), containing admission notes, discharge summaries, operative reports, and consultation notes created between 2004 and 2008.

For non-alphanumeric symbols of interest ('+', '–', '/', and '#'), a target instance of a symbol was defined as the presence of the symbol character within a target token. For the purposes of this pilot, the symbols from institution-specific formatting and various section/headers were excluded. For each symbol, 1,000 instances within the corpus were randomly selected for manual annotation.

Reference standard

Using the General Architecture for Text Engineering (GATE) toolkit(15), each of the 1,000 target symbol instances was marked up within each document to clarify and streamline the process of annotating each target symbol. This was particularly important, as multiple instances of potential symbols may exist within a given text or a given target word token. Although studies have demonstrated that most individuals can interpret the proper meaning of a word with a window size of five,(16, 17) we provided the entire document during annotation of symbols to ensure adequate context.

Our reference standard was created by two annotators with expertise in medicine and linguistics respectively. Because '+' had several medicine-specific meanings, the annotator for this set was a physician. Since meanings of the other four symbols were less medically-specific, a linguist (JR) annotated these samples. Whenever the linguist or physician had questions as to the sense of a symbol, these examples were presented and adjudicated with the assistance of two of the authors with linguistics and medical expertise respectively (SP and GM). "Clinical Corpus Sense" represents this empirically-derived clinical sense inventory for the target symbols. Separately, a second annotator examined 200 random samples (50 per symbol) to establish inter-rater reliability of these annotations with percent agreement and Kappa statistic.

Automated system development and evaluation

We created an initial set of features based on the BoW approach to feature extraction and word-form information within the target and surrounding word tokens. These were compared to the majority sense distribution as the baseline. Three fully supervised classification algorithms were applied to these feature sets in a 10 fold cross-validation setting. These algorithms are Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) implemented with NaïveBayes, LibSVM, and J48 using Weka software(18). We separated 100 random samples from our 1,000 instances of each symbol to determine additional heuristic rules associated with word-form information. After developing the system on 100 random instances, then we evaluated the 900 instances using a 10 fold cross-validation setting on these samples for our result. We report accuracy, recall, precision, and f-measure of our system performance.

Table 1. Senses for symbols.

Symbol	Literature Sense	Reference	Clinical Corpus Sense
+	acid (reaction)	SMAAS, SHC, ICOIEI	
	added to	SMAAS	
	convex lens	SMAAS	
	decreased or diminished (reflexes)	SMAAS	reflexes
	excess	SMAAS	excess
	less than 50% inhibition of hemolysis (Wassermann)	SMAAS	
	low normal (reflexes)	SMAAS	edema(swelling)
	markedly impaired (pulse)	SMAAS	pulse
	mild (pain or severity)	SMAAS	
	plus	SMAAS, SHC, ICOIEI, Kuhn	plus
	positive (laboratory test)	SMAAS, SHC, ICOIEI	positive (laboratory test)
	present	SMAAS	present
	slight reaction or trace (laboratory tests)	SMAAS	
	sluggish (reflexes)	SMAAS	strength
	somewhat diminished (reflexes)	SMAAS	blood type
	and	ICOIEI, Kuhn	and
			pregnancy dating
			heart murmur
			fetal position during labor
			tonsil size
			uncommon rating
–	line-breaking hyphens	FSNLP	line-breaking hyphens
	lexical hyphens	FSNLP	lexical hyphens
	compound pre-modifiers	FSNLP	compound pre-modifier
	quotative or expressing a quantity or rate	FSNLP	quotative or expressing a quantity or rate
	typographic conventions	FSNLP	typographic convention
	phone number	FSNLP, SLP	phone number
	minus	SHC, ICOIEI	minus
			date
			negative
			and
			and(fraction)
			compound
			hyphenated name
			junction
			obstetrical data
			protocol number
			to
			ZIP+4 code
/	divided by	SMAAS	divided by
	either meaning	SMAAS	either meaning
	extension	SMAAS	
	extensors fraction	SMAAS	
	of	SMAAS	of
	per	SMAAS, Kuhn	per
	to	SMAAS	
	date	SLP	date
	separates two doses	Kuhn	separates two doses
			over(e.g., blood pressure)
			abbreviation
			phone number
			respectively
#	fracture	SMAAS	
	gauge	SMAAS	gauge
	number	SMAAS, MT, ICOIEI, FSNLP	number
	pound	SMAAS, MT, ICOIEI	
	weight	SMAAS	
			quantity
			level

SMAAS = Stedman's Medical Abbreviations, Acronyms & Symbols (Forth Edition)
SHC = Stanford Hospital and Clinics approved abbreviations acronyms and symbols
ICOIEI = Illinois College of Optometry and Illinois Eye Institute
Kuhn = Abbreviations and acronyms in healthcare: When shorter isn't sweeter
MT = Medical Terminology the language of health care second edition
SLP = Speech and Language Processing
FSNLP = Foundations of Statistical Natural Language Processing

Table 2. Definition, examples and numbers of symbol senses in clinical documents.

Symbol	Clinical Corpus Sense	Definition	Example	N†
+	pulse	used in pulse degree format	pulses are 2 + bilaterally	287
	edema(swelling)	used in edema degree format	4 + brawny edema	187
	reflexes	used in reflexes degree format	2 + patellar reflexes	148
	pregnancy dating	using in pregnancy dating format	38 + 3 weeks' gestation	115
	excess	more than the given number	20 + years, 37 + weeks	68
	strength	used in strength degree format	strength of the upper extremities is 5+	52
	plus	addition between two numbers	49 + 5 cm	35
	heart murmur	used in heart murmur degree format	there was + 1 mitral regurgitation	23
	blood type	indicates antigen to blood type	a blood type A +	21
	positive (laboratory test)	react to laboratory test	blood pressures with 1-2 + protein	18
	uncommon rating*	uncommon rating	left knee has a 2+ effusion	15
	and	functions like the conjunction *and*	caltrate 600 + vitamin D1	11
	present	exist or react	+/+ trigger points	11
	fetal position during labor	position format during labor	the cervix at + 1 to +2 station	6
	tonsil size	indicates of size of tonsil	3 + tonsils	3
–	quotative, or expressing a quantity or rate	appears in quotatives, or constructions expressing a quantity or rate	5-years-old, once-in-a-lifetime	252
	compound pre-modifier	appears in compound pre-modifiers	seizure-like symptoms	226
	compound	links components of a non-modifier compound	K-Dur, x-ray, E-coli, break-through	157
	lexical hyphen	links small word formatives and content words	non-medically, ex-smoker	126
	to	indicates a range	3-4 times	111
	typographic convention	typographic-conventional hyphen or dash	allergies – none	54
	junction	notes the junction of two elements, usually vertebrae	status post C3-C4 laminectomy	24
	phone number	used in phone-number formatting	612-555-5555	13
	and (fraction)	links an integer and fraction to form a non-integer number	37-1/2 weeks gestation	10
	obstetrical data	appears in what is usually four-pronged data about a patient's pregnancy history	para 0-0-1-0	7
	hyphenated name	links two components of a hyphenated name, usually a surname	Avera-McKennan Hospital	7
	and	functions like the conjunction *and*	type II-III odontoid fracture	3
	date	used in date formatting	05-17-2003	3
	negative	indicates a negative number	-2.132	2
	line-breaking hyphen	follows the first portion of a word that is split by a line break	postopera- tively	2
	ZIP+4 code	separates a zip code and ZIP+4 code	55433-5841	1
	protocol number	serves specification function in an institution's protocol-numbering system	per our protocol #2005-02	1
	minus	indicates subtraction operation	normal 24 + or - 3 ml/kg	1
/	date	used in date formatting	05/17/2003	499
	over(e.g., blood pressure)	couples systolic and diastolic blood pressure measurements, or inhalation and exhalation with BiPAP settings	blood pressure 140/90, we will continue BiPAP at 10/5	196
	either meaning	used in constructions indicating either/both words	and/or, DNR/DNI, Heme/Onc	119
	of	separates a specific rating and the maximum value possible given the scale	regular rate and rhythm with a 2/6 systolic murmur	60
	separates two doses	indicates two separate dosages, usually in drugs with multiple drug constituents	advair 250/50	43
	divided by	separates the numerator and denominator in a fraction	1/2 day, 3-5/7 weeks	39
	per	shorthand for *per*	mg/dL	30
	abbreviation	used to abbreviate, or to link components of an acronym	OB/GYN	6
	respectively	couples values that are each respective to a distinct measure	DP and PT are 1+/4+	6
	phone number	used in phone-number formatting	612/123-4567	2
#	number	shorthand for *number*	hospital day #2	856
	quantity	indicates a quantity, usually of pills dispensed	#10 tablets, #20 dispensed	130
	gauge	indicates gauge specification	aortic valve replacement with #23 medtronic Mosaic valve	13
	level	indicates what level a measurement is at	hemoglobin at #10	1

† N = the number of samples per sense of given symbol in 1000 random samples

* Uncommon rating = subspecialty or other uncommon standard rating

Basic features

- Basic features used as inputs for the three classifiers were:
- Target word token *w* containing the symbol.
- Prefix and postfix of symbol within the targeted word token *w*.
- Previous word tokens *w-n,* target word token *w,* and post one word token *w+n* without stemming (BoW with window size n). We explored the optimal window by varying its size and the effect on performance.

In the example: "….erythema. DTRs are diminished at 1+/4+ in the upper and lower extremities….", if the first '+' symbol (bolded) is the target symbol, the target word token *w* is "1+/4+", the prefix is "1", the postfix is "/4+". BoW with window size 1 is {at, 1+/4+, in} and BoW with window size 2 is {diminished, at, 1+/4+, in, the}.

Beside basic features, we experimented with stop word removal with BoW using a standard list of 57 English stop words(8). With our previous example, stop word removal with BoW window size 1 is {diminished, 1+/4+, upper}, and the set of BoW with window size 2 and without stop words is {DTRs, diminished, 1+/4+, upper, lower}.

Heuristic features

We tested heuristic rules as additional features. Heuristic rules were developed to identify word-form representations of the target word token w or surrounding word tokens (w-n and w+n). 100 random instances from 1,000 were separated for each symbol to develop heuristic rules. Utilizing regular expressions, heuristic rules applied to the target word token w or surrounding word tokens. These were added as additional features to our classifiers.

Results

Table 1 compares literature senses from reference sources and the experimental clinical corpus senses in our repository for each symbol. This comparison is organized in the alphabetical order of literature senses. Depending upon the domain, a different set of senses was identified. Table 2 depicts the sense, its definition, an example, and the distribution of senses within the corpus. Table 2 is ordered based on the sense distribution of each symbol within the clinical corpus. When developing our module, we introduced heuristic rules for this pilot as depicted in Table 3.

Table 3. Heuristic rules used as additional features to classifier.

Symbol	Regular expression	Description of form	Applied sense
+	m/^[1-3]\+/	1+, 2+, 3+	pulse, edema, reflex, excess
	m/^\+[1-3]/	+1, +2, +3	pulse, edema, reflex, excess
	m/^[1-9]?[0-9]\+[0-9]\W?$/	one or two digits for weeks with both side of '+'	pregnancy dating
	m/^[1-9][0-9]\+\W?$/	two digits for years/weeks with previous side of '+'	excess
–	m/^[1-9][0-9]\-/	two digits with previous side of '–'	
	m/^.+\-[1-9][0-9]$/	two digits with post side of '–'	
	m/^[a-zA-Z]?\-[a-zA-Z]?$/	two alphabetic words with both side of '–'	compound, lexical hyphen
	m/^[a-zA-Z]?\-[a-zA-Z]\-[a-zA-Z]?$/	three alphabetic words with both side of two '–'	quotative
/	m/^(1[0-9][0-9])\/((1?[0-9][0-9])\W?)$/)	two or three digits with both side of '/'	over(e.g., blood pressure)
	m/^([a-zA-Z]+)\/([a-zA-Z]+)\W?$/	two alphabetic words with both side of '/'	either meaning
	m/^[0-9]\/([0-9])\W?$/	two digits with both side of '/'	of
#	m/^\#1[0-5]\W*\.*\W*$/ or m/^\#[1-9]\W*\.*\W*$/)	one or two digits for days with post side of '#'	number
	m/^\#[1-4][0|5]\W*\.*\W*$/	two digits for quantity with post side of '#'	quantity

Within the overall corpus of 604,944 notes, the frequency of '+', '–', '/', and '#' are represented in Table 4. For inter-rater reliability, 50 random samples were annotated by a second annotator. Proportion agreement and Kappa statistic of each symbol in Table 4 indicates respectively reasonable inter-rater agreement even if it is conducted in a small size of samples.

Table 4. Frequency in total corpus and inter-rate agreement of symbols

Symbol	Frequency	Proportion agreement (%)	Kappa statistic
+	118,283	100	1.00
–	4,821,029	88	0.86
/	4,785,691	96	0.95
#	721,655	90	0.72

When we applied three supervised machine-learning algorithms with our feature sets, NB classifier had the most stable overall performance compared to both SVM and DT classifier. We tested removal of stop words; however, there was no performance improvement. We also added heuristic rules as described in Table 3, but there is no significant change in algorithm performance either. Our results with respect to the accuracy, recall, precision, and f-measure with the NB, SVM, and DT classifiers using the basic feature set alone and with BoW are summarized in Table 5. These results are based on 900 test samples considering all separated senses in Table 2. Maximum accuracy for symbol '+' was 80.11%, symbol '–' - 80.22%, symbol '/' - 90.44%, and symbol '#' - 95.00% with the NB classifier. For '+' and '/', using BoW as features provided improved performance with the NB classifier (Table 5), but the optimal window size was different for each symbol. For '–', BoW did not contribute additional information for symbol disambiguation. For '#', the target symbol alone was the dominant feature of importance.

Table 5. Performance of Naïve Bayes, Support Vector Machine, and Decision Tree classifiers. Acc = Accuracy, Pre = Precision, Sen = Sensitivity, F-m = F-measure

Symbol	Feature	Naïve Bayes				Support Vector Machine				Decision Tree			
		Acc*	Pre*	Sen*	F-m*	Acc*	Pre*	Sen*	F-m*	Acc*	Pre*	Sen*	F-m*
+	Majority	0.29	0.08	0.29	0.13	0.29	0.08	0.29	0.13	0.29	0.08	0.29	0.13
	Target token	0.47	0.52	0.47	0.41	0.52	0.47	0.52	0.47	0.49	0.49	0.49	0.45
	Target token, Prefix/postfix	0.54	0.51	0.54	0.48	0.54	0.50	0.54	0.48	0.49	0.49	0.49	0.45
	Target token, BoW (size = 1)	0.68	0.66	0.68	0.63	0.41	0.56	0.41	0.36	0.65	0.67	0.65	0.62
	Target token, BoW (size = 2)	0.77	0.74	0.77	0.74	0.32	0.57	0.32	0.20	0.66	0.67	0.66	0.63
	Target token, BoW (size = 3)	0.79	0.75	0.79	0.76	0.30	0.37	0.30	0.15	0.65	0.67	0.65	0.63
	Target token, BoW (size = 4)	**0.80**	**0.78**	**0.80**	**0.78**	0.29	0.22	0.29	0.13	0.65	0.67	0.65	0.62
	Target token, BoW (size = 5)	0.80	0.78	0.80	0.78	0.29	0.22	0.29	0.13	0.65	0.67	0.65	0.63
–	Majority	0.25	0.06	0.25	0.10	0.25	0.06	0.25	0.10	0.25	0.06	0.25	0.10
	Target token	0.62	0.73	0.62	0.61	0.63	0.68	0.63	0.62	0.63	0.79	0.63	0.63
	Target token, Prefix/postfix	**0.80**	**0.80**	**0.80**	**0.79**	0.66	0.79	0.66	0.67	0.63	0.79	0.63	0.63
	Target token, BoW (size = 1), Prefix/postfix	0.77	0.77	0.77	0.76	0.32	0.65	0.32	0.24	0.63	0.79	0.63	0.63
/	Majority	0.51	0.26	0.51	0.34	0.51	0.26	0.51	0.34	0.51	0.26	0.51	0.34
	Target token	0.57	0.49	0.57	0.46	0.61	0.64	0.61	0.55	0.51	0.26	0.51	0.34
	Target token, Prefix/postfix	0.84	0.86	0.84	0.83	0.64	0.75	0.64	0.57	0.75	0.81	0.75	0.71
	Target token, BoW (size = 1), Prefix/postfix	**0.90**	**0.89**	**0.90**	**0.90**	0.54	0.60	0.54	0.40	0.72	0.66	0.72	0.62
	Target token, BoW (size = 2), Prefix/postfix	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.72	0.66	0.72	0.66
#	Majority	0.85	0.72	0.85	0.78	0.85	0.72	0.85	0.78	0.85	0.72	0.85	0.78
	Target token	**0.95**	**0.94**	**0.95**	**0.94**	0.94	0.93	0.94	0.93	0.95	0.94	0.95	0.94
	Target token, Prefix/postfix	0.92	0.92	0.92	0.92	0.94	0.92	0.94	0.93	**0.96**	**0.94**	**0.96**	**0.95**
	Target token, BoW (size = 1)	0.94	0.95	0.94	0.95	0.86	0.86	0.86	0.80	0.95	0.95	0.95	0.94

Discussion

We examined non-alphanumeric symbol disambiguation, an under-studied pre-processing NLP function in the clinical domain. To gain a more thorough understanding of symbol sense ambiguity, we performed a survey of the literature and generated an empiric sense inventory, which helped to refine the overall inventory. Symbol disambiguation appears to perform well with simple sets of features but requires different combinations of features for individual symbols. In each case, a relatively small set of features based on the symbol and its context were effective, indicating that this is a relatively simpler task than sense disambiguation for words, acronyms and abbreviations. Despite the relative simplicity of the task, it has been largely ignored in the clinical NLP literature but constitutes an important problem for NLP of clinical documentation. For example, being able to determine the context appropriate meanings of symbols can contribute to improved named entity recognition and classification.

While the surrounding context, including words beyond the target token, were expected to be important, we found that in the cases of '#' and '–', words beyond the target word *w* were unnecessary. In fact, for '#', the target word *w* alone was sufficient for excellent performance. In contrast, senses related to '+' required surrounding context (optimized with window size 4) for optimal performance. One of the main reasons for these differences is that symbol resolution is affected by the number of senses in the sense inventory and proportion of the majority sense of each symbol. In the previous example, the '#' symbol has fewer senses and has higher propotion of the dominant sense (only 4 senses and the majority sense prevalence is 85%) compared to '+' symbol with has 15 possible senses with well-balanced distributions. For '–' symbol, it only required isolated and condensed token information (pre and postfix features) to determine the right meaning. Another potential reason is the degree of semantic relatedenss

among senses in a given symbol. For example, '#' symbol has 4 senses that are all closely related with the concept 'number'. Thus, disambiguation of the '#' symbol results in better performance compared to '–' symbol, which had a variety of concepts such as 'minus' or several lexical expressions (e.g., lexical hyphens, compound pre-modifier).

We also expected heuristic rules to contribute positively to system performance but found that they were not helpful in our experiment. These rules could be helpful for enumerated items such as dates or telephone numbers where training sets may not be sufficient to capture the large number of possible combinations. We speculate that this did not change performance much since both of these items were low-incidence. Also, some rules are language and format-specific. For example, the form of date with symbol '–' can be different according to location. The sequence of date, month, and year are opposite in Europe/Asia compared with the United States. Because of these limitations and perhaps some overlap with our general form-based features (thereby not being independent of our heuristic rules), heuristic features did not contribute significantly to system performance.

With the '+' symbol, we discovered that there were a number of senses that were specific to subspecialties or occurred less often, which we combined into a single annotation called "uncommon rating". In contrast, common ratings such as that for edema or reflexes were separated out as separate senses. For example, the sense 'effusion of a joint' (e.g., "Left knee has a 2+ effusion…") or 'prostate size' (e.g., "His is prostate is 1 to 2+ enlarged…") are standard but occur with low frequency. If we group less common senses into a single annotation, the performance of automatic symbol resolution module improves. In this study, we grouped these less common senses into one sense. If we extend this, all kinds of senses such as pulse, strength, reflexes, edema, and uncommon ratings for symbol '+' can be grouped together. As expected, with this aggregate set of senses, the accuracy of NB classifier was 88.89%, up from 81.56% when more common ratings were separated from the less common ratings. Because these sense grouping decisions can be somewhat arbitrary or tailored to the purpose of the NLP module, concrete agreement between annotators and a clear understanding of the goals of the particular symbol disambiguation NLP module's scope are essential.

Another issue is that some senses share the same BoW or the same form of the target token. In the symbol '–' set, for example, "follow-up", "well-nourished" and "seizure-like" can be a lexical hyphen and/or a compound-premodifier. For example, "He arrived on time for his follow-up" and "His symptoms were seizure-like", these '–' instances are categorized as lexical hyphens, while "We scheduled his follow-up appointment" and "He experienced seizure-like symptoms" are considered to be both lexical hyphens and compound pre-modifier hyphens. These shared forms between senses may create difficulties with disambiguation and may require additional syntactic information such as part-of-speech and syntactic phrase category. The distinction between lexical hyphens and compound pre-modifier hyphens is probably too small to be of practical importance in an NLP system; however, in this exploratory study we chose separate annotations for these entities that may be collapsed.

Our research demonstrates that non-alphanumeric symbol disambiguation is feasible, with good performance on clinical text using standard form-based rules. These rules require some calibration for each symbol type with respect to window size for individual symbols. Since the set of non-alphanumeric symbols is finite (vs. words and acronyms), development of fully supervised disambiguation classifiers is likely to be the most effective and accurate approach. We plan to extend this module to other symbols, including alphabetic symbols, such as 'x', as well as additional non-alphanumeric symbols, with the goal of utilizing these techniques within a pre-processing module for down-stream information extraction functions from clinical text.

Conclusion

Although symbols, primarily non-alphanumeric characters, are used widely to convey a variety of meanings in clinical discourse, symbol resolution has been less studied by the linguistics and medical NLP communities. Symbol resolution can be viewed as a specific type of WSD, as well as a basic module for automatic medical NLP systems. In this paper, we examined four symbols ('+', '–', '/', and '#') to detect clinical symbol senses and to contrast with senses attested in the literature. We found that while supervised machine learning approaches with form-based features to be effective, calibration of features for disambiguation may be needed for system optimization with individual symbols.

Acknowledgements

This research was supported by the American Surgical Association Foundation Fellowship, the University of Minnesota Institute for Health Informatics Seed Grant, and by the National Library of Medicine (#R01 LM009623-01). We would like to thank Fairview Health Services for ongoing support of this research.

References

1. Stetson PD, Johnson SB, Scotch M, Hripcsak G. The sublanguage of cross-coverage. Proc AMIA Symp. 2002:742-6.
2. Watson R, editor. Part-of-speech Tagging Models for Parsing. Pro of the 9th Annual Computational Linguistics community in the UK Colloquium; 2006; Open University, Milton Keynes.
3. Yoshida K, Tsuruoka Y, Miyao Y, Tsujii Ji. Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers. Proceedings of the 20th international joint conference on Artifical intelligence; Hyderabad, India. 1625564: Morgan Kaufmann Publishers Inc.; 2007. p. 1783-8.
4. Dell'orletta F, editor. Ensemble system for part-of-speech tagging. Proc of the 11th Conference of the Italian Association for Artificial Intelligence; 2009; Reggio Emilia, Italy.
5. Xu H, Fan J-W, Hripcsak G, Mendonca EA, Markatou M, Friedman C. Gene symbol disambiguation using knowledge-based profiles. Bioinformatics. 2007 April 15, 2007;23(8):1015-22.
6. Hwang FL, Yu MS, Wu MJ, editors. The improving techniques for disambiguating non-alphabet sense categories. Proc of Research on Computational Linguistics Conference XIII; 2000.
7. Yu MS, Hwang FL. Disambiguating the senses of non-text symbols for Mandarin TTS systems with a three-layer classifier. Speech Communication. 2003;39(3-4):191-229.
8. Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press; 1999.
9. Jurafsky D, Martin JH. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J.: Prentice Hall; 2000.
10. Stedman's Medical Abbreviations, Acronyms & Symbols. 4 ed2008.
11. Willis MC. Medical Terminology: The Language of Health Care. 2 ed2006.
12. SHC approved abbreviations acronyms and symbols. Stanford hospital and clinics; Available from: http://stanfordhospital.org/forPhysiciansOthers/physicians/documents/SHCApprovedAbbreviations.pdf.
13. Approved and unapporved abbreviations and symbols for medical records. Illinois college of optometry and Illinois eye institute; 2009; Available from: http://www.ico.edu/policies/Information%20Management/Approved%20and%20Unapproved%20Abbreviations%20for%20Medical%20Records.pdf.
14. Kuhn IF. Abbreviations and acronyms in healthcare: when shorter isn't sweeter. Pediatr Nurs. 2007 Sep-Oct;33(5):392-8.
15. Cunningham H, Maynard D, Bontcheva K, Tablan V, editors. GATE: A framework and graphical development environment for robust NLP tools and applications. Proc of the 40th Anniversary Meeting of the Association for Computational Linguistics; 2002.
16. Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. J Comput Biol. 2005 Jun;12(5):554-65.
17. Kaplan A. An experimental study of ambiguity and context. Mechanical Translation. 1950;2(2):39-46.
18. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009;11(1):10-8.