journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Assessment of commercial NLP engines for medication information extraction from dictated clinical notes

V. Jagannathan[b,*], Charles J. Mullett[a], James G. Arbogast[a], Kevin A. Halbritter[a], Deepthi Yellapragada[a], Sushmitha Regulapati[a], Pavani Bandaru[a]

[a] West Virginia University, United States
[b] MedQuist Inc., 235 High Street, Suite 213, Morgantown, WV 26505, United States

ARTICLE INFO

Article history:
Received 26 September 2007
Received in revised form
15 August 2008
Accepted 19 August 2008

Keywords:
Natural language processing (NLP)
Medication extraction
Text mining

ABSTRACT

Purpose: We assessed the current state of commercial natural language processing (NLP) engines for their ability to extract medication information from textual clinical documents.
Methods: Two thousand de-identified discharge summaries and family practice notes were submitted to four commercial NLP engines with the request to extract all medication information. The four sets of returned results were combined to create a comparison standard which was validated against a manual, physician-derived gold standard created from a subset of 100 reports. Once validated, the individual vendor results for medication names, strengths, route, and frequency were compared against this automated standard with precision, recall, and *F* measures calculated.
Results: Compared with the manual, physician-derived gold standard, the automated standard was successful at accurately capturing medication names (*F* measure = 93.2%), but performed less well with strength (85.3%) and route (80.3%), and relatively poorly with dosing frequency (48.3%). Moderate variability was seen in the strengths of the four vendors. The vendors performed better with the structured discharge summaries than with the clinic notes in an analysis comparing the two document types.
Conclusion: Although automated extraction may serve as the foundation for a manual review process, it is not ready to automate medication lists without human intervention.

1. Introduction

Natural language processing (NLP) technology has a long history in computer science and is an active area of research in healthcare. By extracting data and information from dictated medical reports, NLP can potentially be used to facilitate and improve the process of medical care—particularly when coupled with contemporary clinical information systems. Early efforts by Sager and Friedman laid the groundwork for the field [1–3]. Since then, academic investigators have focused on extraction of findings and problems [3,4], while the healthcare industry has described commercial business products [5–8]. Computer assisted coding for the submission of billing claims is a thriving market [9–12]. More recently, NLP-focused research challenges have shined a spotlight on the role of NLP in medicine [13,14].

Relatively high rates of medication-related errors associated with the transition of care environments led the Joint Commission for Accreditation of Healthcare Organizations (JCAHO) to mandate a process for medication reconciliation across the continuum of care. This mandate charges physicians with explicitly verifying and continuing or discon-

* *Corresponding author*. Tel.: +1 304 296 7550x101.
E-mail address: juggy@medquist.com (V. Jagannathan).
1386-5056/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2008.08.006

tinuing each medication on the patient medication list when admitting, transferring, or discharging between and among healthcare facilities and the home environment [15]. Medication reconciliation and medication list management have therefore been targeted for automation by recent investigators who developed their own specialty applications to pull and merge coded and narrated sources of medication information from their local records for display to the users [16,17]. We hypothesized that *commercial* NLP technologies had reached the point in their development that they were ready to create medication lists from dictated clinic notes and discharge summaries.

We therefore designed an analysis that compared medication extraction from four different vendors against a benchmark standard developed using the results of all four products. We validated this automated benchmark with a manual, physician-derived gold standard from a subset of 100 of the reports. Our analysis mimics the method described by Turchin et al. [18], but with a larger corpus and a focus on practicality and deployable workflow solutions.

2. Methods

2.1. Creating the corpus

We obtained IRB approval for a corpora of 2000 dictated reports on the condition that automated de-identification would be successfully confirmed by hand-review of a random sample of 50. The target metric for confirmation of a successful de-identification method was that no more than 5 of the 50 reports would contain any residual element of protected health information (PHI).

Following approval by the West Virginia University Institutional Review Board, we obtained 1000 discharge summaries and 1000 family practice clinic notes. Patient names and other protected health information were removed using the commercially available de-identification software, De-Id [19]. The process steps used to create the corpus are summarized below:

- Clinical documents were obtained using an HL7 interface.
- The messages were processed and reformatted to a format acceptable to the de-identification software from De-Id Inc.
- The text portion of the data was first de-identified using De-Id software.
- The header data was then de-identified while maintaining relationships that may exist between multiple documents.
- The success of the de-identification process was confirmed by a Health Information Management officer of the hospital who found only one instance of partial PHI – a patient's first name – in the sample of 50 documents.

2.2. Evaluation approach

We distributed these 2000 reports to three vendors and asked them to extract the medication-related information and return to us. A fourth vendor was added later in the process.

We then performed two sequential evaluations, one to gauge the success of our automatically generated standard and the second to measure each vendor's success against this standard. For the first evaluation, we picked a random selection of 50 discharge summaries and 50 family practice notes with between 3 and 30 medication mentions in the first vendor's output. This collection of 100 documents was then evaluated by our physician reviewer who extracted the medication information by hand to create the gold standard. The four vendor result sets were combined to generate an automated comparison-standard through a process detailed further below. The success of this automated comparison standard was evaluated by comparing its results to the physician gold standard. Once the automated comparison-standard was created and validated, we evaluated the success of individual vendors against the benchmark of the automated comparison standard for the full set of 2000 reports. Management of unanticipated difficulties arising secondary to non-standard medication vocabularies and other differences between the vendors required manual manipulations detailed below.

2.3. Vendors and medication extraction

The de-identified reports were sent to four different NLP vendors (Language and Computing, Coderyte, LingoLogics and Artificial Medical Intelligence) with instructions to extract medication information. Very little detailed technical information is available on these commercial and proprietary solutions. A brief description from what is publicly disclosed by the vendors is provided below:

- *Language and Computing*—a Belgian company with one of the longest tenures in commercial NLP-based solutions. Their approach relies on an indigenously developed ontological taxonomy of clinical concepts. They have focused on concept abstraction with mapping to SNOMED. Recently, they have also adapted their approach to support ICD9 coding and billing.
- *Coderyte*—a recent entry into the NLP marketplace. They take a statistical approach to their NLP tasks and their primary focus is coding of ambulatory clinical notes for assignment of ICD9, CPT and E&M levels associated with billing.
- *LingoLogics* – a recent entry into the NLP marketplace – markets coding solutions. The genesis of their commercial solutions is based on research work done at the Mayo Clinic and relies on an ontological approach. Their commercial focus is the coding of ambulatory clinical notes with ICD9, CPT and E&M levels.
- *Artificial Medical Intelligence (AMI)*—also a recent entry into the NLP marketplace. They have developed a proprietary indigenous hybrid approach combining terminology and rules in their product. The commercial focus in conjunction with 3M encoders is to support coding and billing in inpatient settings.

Each NLP vendor returned the data set with their own proprietary XML-markup of the data. Special interface routines were developed to process these markups to extract the concepts being evaluated. In the future, one would anticipate that

vendors would use the HL7 Clinical Document Architecture (CDA) standard to represent extracted concepts and structure [22], making it easier for the receiver to interpret the results. The extracted medication information was then stored in a database table as XML-snippets. This result set was used to perform the analysis.

There are eight dimensions which are relevant for medication extraction:

(1) Medication name (drug mentioned in the report);
(2) Medication form (pill, liquid, powder, etc.);
(3) Dosage (number);
(4) Medication strength (strength and unit);
(5) Route of administration (oral, intravenous, etc.);
(6) Frequency (how often administered);
(7) Duration (period of time and duration units);
(8) Medication status (active/inactive).

Of the four NLP vendor solutions that we tested, only one vendor (Language and Computing) attempted to extract all of these eight elements. To make meaningful comparison across these vendors we decided to focus on the four dimensions shown below, which were extracted by the majority of the NLP systems:

(1) Medication name (drug mentioned in the report);
(2) Medication strength (strength and unit);
(3) Route of administration (oral, intravenous, etc.);
(4) Frequency (how often administered).

2.4. *Development and evaluation of the comparison standard*

Although gold standards are typically created using manual review, we postulated that it was feasible to do an effective evaluation by leveraging the result set from multiple vendors. We therefore devised an automated process that used the four vendors to create the comparison standard. Essentially, agreement between any two vendor solutions on medication name was added to the target gold medication list. Once a match on the medication name was confirmed, the other dimensions abstracted were examined to determine consensus values for them. Table 1 highlights a sample extraction from two vendors which illustrates the problem of mismatches secondary to a lack of standardization of medication name. For the purpose of our evaluation, we assumed that if any string was a proper subset of another, the medications matched. We also ignored any punctuation in the process such as "–" in the Senokot example in the table.

Table 1 – Examples of variances in medication names extracted by two different vendors

Document #	Vendor X	Vendor Y
644	Percocet tablets	Percocet
	Vitamin multi	Vitamin
	Senokot S	Senokot-S
	Dulcolax	Dulcolax suppository
	Tylenol tablets	Tylenol
	Intrathecal baclofen	Baclofen
	Antacid choice	–
736	Imdur ER	Imdur
	Renagel	–
	Antimicrobial	–
886	Gentamicin cream	Gentamicin
	Prilosec counter	Prilosec
	Multivitamin	–

We analyzed the success of the establishment of the automated comparison standard by comparing it to the manual, physician-derived gold standard for the 100 validation reports, calculating precision, recall, and the *F* measure. The formulae used are the following:

$$\text{precision } P = \frac{TP}{TP + FP}$$

$$\text{recall } R = \frac{TP}{TP + FN}$$

$$F \text{ measure} = \frac{2PR}{P + R}$$

where TP is true positives, FP is false positives and FN is false negatives.

We manually assisted the matching of the results obtained on other dimensions of medication-related information—dose, route, and administration frequency as well. For instance, medication administration frequency was not handled uniformly between the different vendors. One vendor attempted to standardize the vocabulary of the frequency, while another captured and reported the exact verbiage of the physician. The result was a wide disparity on extraction of the same concept. For example, the concept of "once daily" was represented many different ways in the dictated and extracted text:

> Once daily, 1times everyday, every day, q. 24h, q. 24 hours, every 24 hours, once a day, per day, 1 time a day, daily

To manage this problem, we created a dictionary lookup table used to manually create equivalence in concepts.

The bulk of the labor of this evaluation occurred in the gold standard creation phase. Once this was established, the evaluations between vendors were more straightforward, tallying each result for each of the 2000 documents against the group result found in the comparison standard. For each vendor, we calculated precision, recall, and the *F* measure for medication name, dose, route and frequency.

3. Results

3.1. *Automated comparison standard versus manual gold standard*

Table 2 shows the result of comparing the standard that was automatically created with that of the gold standard of physician manual review. The automated comparison standard was most accurate at recognizing medication names, with precision, recall, and *F* measure scores above 90%. This comparison standard also scored reasonably well with medication

Table 2 – Baseline comparison of automated gold standard with manual gold standard on a sample size of 100 reports

Dimension	Name (%)	Strength (%)	Frequency (%)	Route (%)
Precision	95.9 (±1.79)	97.9 (±1.84)	71.3 (±6.72)	98.0 (±2.74)
Recall	90.6 (±2.53)	75.6 (±4.84)	36.8 (±5.13)	68.1 (±7.62)
F measure	93.2	85.3	48.3	80.3

95% confidence intervals are given in parentheses.

strength and route, but did poorly for frequency. In light of this deficiency, we omitted individual vendor comparisons on frequency.

3.2. *Individual vendors versus automated comparison standard*

Table 3 shows the performance of individual vendors against the automated comparison standard for the entire set of 2000 reports. Essentially, the vendors did reasonably and uniformly well with identifying medication names, dosing strengths and route, with only minor variations in relative strengths among the metrics. Vendors 1 and 2 had similarly high scores with medication names, whereas vendors 3 and 4 had difficulty with medication name recall. Vendor 4 scored well with dosage strength, but did not attempt to analyze the route. Vendor 3 did the best at route precision, but did not perform as well as 1 and 2 at recall. Overall, vendor 1 scored the highest, with an *F* measure mean of 93%, whereas vendors 2, 3, and 4 scored 91%, 87%, and 81%, respectively.

In a sub-analysis comparing the results against the clinic notes versus the results against the discharge summaries, we found that the vendors performed better with the discharge summaries (four-vendor *F* measure means for discharge summaries = 90.5%; clinic notes = 78.7%). The fact that vendors did better with discharge summaries is not unexpected, as the medications mentioned in discharge summaries are typically more structured. Of the three elements (name, strength, and route), route was by far the biggest challenge in the clinic notes, with a four-vendor *F* measure mean of 68% versus 84.7% for names and 80.4% for strengths.

4. Discussion

We hypothesized that commercial NLP vendors operating in the healthcare domain had progressed the development of their products to an extent that they were ready to extract medication information to populate a list of active patient medications. While the medication name identification was reasonably well performed, the context of the medication mention was absent, and therefore our initial hypothesis was too optimistic. This missing context leads to a loss of precision, as medications are falsely identified for inclusion on the medication list when they were mentioned for other reasons such as describing past use, or indicating a past allergic reaction. Recall for medication names, i.e. capturing all the current medications mentioned, is an easier task for the NLP engines, as long as a robust medication dictionary is embedded. Our results found vendors 1 and 2 were especially strong with recall of medication names with percentage scores in the high 90 s. However, vendor 4 found only 52% of the medication names, but conversely had the highest precision, scoring 100%.

In terms of the automated building of a gold standard, our approach appears feasible under two conditions: (a) the initial recognition of concepts is reasonably high and (b) one has access to multiple comparison solution sets developed independently of each other, i.e. from competing vendors. In our example, both conditions were satisfied for medication names, medication strength, and route. However, it is likely that the automated method included instances of medication mentions in the comparison standard that would not have been placed on an active medication list by our physician,

Table 3 – Vendor performance on 2000 reports for (a) medication name extraction, (b) medication strength extraction, (c) medication route extraction

	V1	V2	V3	V4
(a) Name				
Precision (%)	98.3 (±0.25)	99.2 (±0.17)	99.6 (±0.13)	100 (±0.05)
Recall (%)	98.2 (±0.25)	97.4 (±0.30)	87.0 (±0.64)	52.0 (±0.96)
F measure (%)	98.2	98.3	92.9	68.4
(b) Strength				
Precision (%)	88.4 (±0.84)	83.7 (±0.91)	72.6 (±1.07)	96.9 (±0.50)
Recall (%)	91.3 (±0.75)	97.9 (±0.38)	89.6 (±0.82)	82.0 (±1.03)
F measure (%)	89.9	90.2	80.2	88.9
(c) Route				
Precision (%)	86.1 (±1.32)	71.5 (±1.56)	95.3 (±0.94)	–
Recall (%)	97.8 (±0.60)	99.2 (±0.36)	79.6 (±1.64)	–
F measure (%)	91.6	83.1	86.7	–

95% confidence intervals are given in parentheses.

such as when a medication is described as being discontinued, or being mentioned as an allergy. This would explain the 100% score from vendor 4 against the automated standard for precision. This NLP engine would not likely have scored perfectly against the physician's gold standard. Indeed, the other vendors' precision scores are likely similarly inflated, due to some actually false positive medication identifications being included as true positives by the automated gold standard generation method. This is an important limitation of this evaluation.

A review of some of our particularly instructive qualitative observations of problem areas and potential solutions follows below.

4.1. *Discussion of problem areas with automated extraction*

The following examples highlight some of the deficiencies and challenges faced by the NLP engines with real-world physician dictations.

4.1.1. *Issues: medication vocabulary*

4.1.1.1. Medication or dietary formula. Is infant formula a medication? One of the vendors confused these terms. "REVIEW OF SYSTEMS: Using Alimentum formula 16 ounces daily. Systems outside of GI and musculoskeletal are negative to direct questioning."

In this case, a vendor extracted *Alimentum* as a medication.

4.1.1.2. Recognition of combination medications. The systems had difficulty extracting medication information when multiple medications are combined into one drug formulation, as in the example, "Emtricitabine/tenofovir 200/300 mg 1 p.o. daily." From this sentence, one vendor identified emtricitabine at one oral dose daily, but did not capture the milligram strength, and identified tenofovir at the erroneous dose of 200 mg administered once daily. In another example, "Atacand/HCTZ 16/12.5 mg p.o. daily," A vendor did not capture a dose, frequency or route for the Atacand, and garbled the dose as "16/12.5 mg" for the hydrochlorothiazide (HCTZ).

In the case of both of these fixed combination drugs, an improved medication dictionary listing such combinations should help the NLP engines' recognition accuracy. Alternatively, the engines could use regular expressions that recognize the individual components of the medication combinations and parse them intelligently.

4.1.2. *Issues: context of medication mention*

4.1.2.1. Medications mentioned as allergies rather than therapies. As an example of one dictation, "ALLERGIES: Nifedipine." Vendors typically extracted these medication names, but lacked context of its mention as an allergy, rather than a medication for the patient.

4.1.2.2. Mom's medication abstracted on baby's note. "On **DATE[Jul 12 2007], afternoon transferred here and admitted in the pediatric floor. A surgical consult was done and she was started on clindamycin, vancomycin IV based on the history of MRSA positive in mother which is a nurse health care provider in **DATE[Jan] when she was working with a MRSA patient. Mom was treated with Bactrim as MRSA was sensitive to that." Here again, the medication "Bactrim" was captured and listed, but without indication of the context—use by the mother of the case patient.

4.1.2.3. Time scale. "Prior to discharge night, the patient had 7.5 mg of Coumadin and the patient was advised to take 5 mg of Coumadin for 2 days and then 4 mg of Coumadin for the next 2 days." In this example, one vendor just listed Coumadin twice, at doses of 5 and 4 mg. The context around the sequence and timing of the decrease in dose was lost.

To properly understand the context of a medication mentioned, an advanced NLP engine requires an additional layer of algorithms or routines designed to capture contextual subtleties like temporal factors (happening now, in the past, to happen in the future), referent (patient or relative), and communication of past allergic reactions.

4.1.3. *Issues: dosing changes based on patient response*

"It is okay to go to 1 clonidine once a day but I told her if the systolic is over 135 on a regular basis she needs to go back to 2 a day." Vendors captured the clonidine mention, but one listed two dosing frequencies, "once a day |2 a day," and the other noted only the first frequency of "one daily," missing the dosing change instructions entirely. It is interesting to note the ambiguity in the physician's dictation; he or she could have meant two tablets once daily or one tablet twice daily. Clonidine is typically dosed at one tablet twice daily, so the first vendor's interpretation was likely on the correct track.

Recognizing dosing changes based on patient condition is probably the hardest for an automated NLP engine to capture. An NLP engine must parse, capture, and logically reconstruct the entire context to interpret these sentences correctly.

4.1.4. *Issues: unexpected numerical data*

Physicians occasionally choose to dictate the concentration of the liquid preparation to be taken by the patient, as in this example, "Augmentin 500 mg (400 mg/5 mL) 6 mL p.o. b.i.d. for 10 days." In this instance, one vendor extracted the medication name, route, and frequency correctly, but gave the dose as "500 mg|400 mg|5 mL|6 mL." The context of the suspension concentration provided by the parentheses and the divisor symbol are lost. A more robust NLP engine would be needed to analyze, capture and present the concentration strength of this medication.

4.2. *Future directions*

From our qualitative examination of the errors made by the vendors, we compiled three suggestions for future research areas to improve performance of medication identification from clinical records: (1) a more robust medication dictionary; (2) a more sophisticated representation of medications; and (3) algorithms for identifying contextual information. These suggestions are further explored below.

4.2.1. *Medication dictionary*

In our qualitative examination of the errors made by the vendors, including the examples shown above, their need for a more robust medication dictionary is highlighted. For

instance, vendor 4 had a very high precision but a low recall on this dimension. In the case of medication names, we feel it is a simply a matter of having better medication dictionary. This is a relatively easy issue to fix.

4.2.2. *Medication representation*

In the case of dosage, frequency or route, some of the problems that the vendors have on this front is representational—particularly how to represent complex concepts such as tapered dosages, conditional use of medications, etc. More experience and evolution of their NLP engines will be needed to improve these extractions.

4.2.3. *Medication context*

The greater challenge is the explicit capture of the larger context of the medication mention in the physician note, for instance, was the medication given to the patient rather than to someone else? Is the medication name mention actually a current therapeutic or is it an allergy or other type of mention? Context recognition, like the work by Chapman [20,21] focuses on using simple regular expression patterns to detect negation, temporal status, whether a condition is hypothetical, and whether the patient experienced the condition. However, this field is not well developed, and determining whether a medication is even active or not remains problematic. The field is even further away from successfully parsing the fact that the Bactrim in the example cited above was being taken by the mother of the infant patient. In our analysis of four current vendors, only one even attempted to determine the context of an active versus inactive status of the medication. We were therefore unable to create a comparison standard for lack of additional vendors' result sets. Given that many of the examples cited above directly or indirectly involve the context of the medication mentions, we feel this global deficiency limits the applicability of these NLP engines.

Our results and experience suggest that any extraction done automatically needs to be vetted by a human. We would therefore advise that parties use these products to highlight in a block of dictated text the names, doses, routes, and frequencies of medications mentioned and then have an editor, perhaps the transcriptionist, determine and sort which ones should be added to the medication list, which should be removed, etc. After reaching this conclusion, we developed a prototype interface to help with this process, shown in Fig. 1. Transcription services companies such as MedQuist may elect to use software such as described to train its transcriptionists to provide medication list management for its clients.

4.3. *Limitations*

Automating the generation of the comparison standard may have allowed the inclusion of medications that a physician reviewer would have excluded due to additional contextual information not considered by the panel of vendor NLP engines. While this may have potentially increased the precision results reported in our individual vendor analysis, it does not influence our general findings and conclusions. We might also have compared the different result sets obtained when increasing the medication name matching inclusion criteria for the automated comparison standard from two of the four vendors to three of the four vendors. By raising this threshold, we would have reduced the number of medications in the comparison standard which would have influenced the results seen in the evaluation of our automated standard versus the physician-derived gold standard, as well as the results

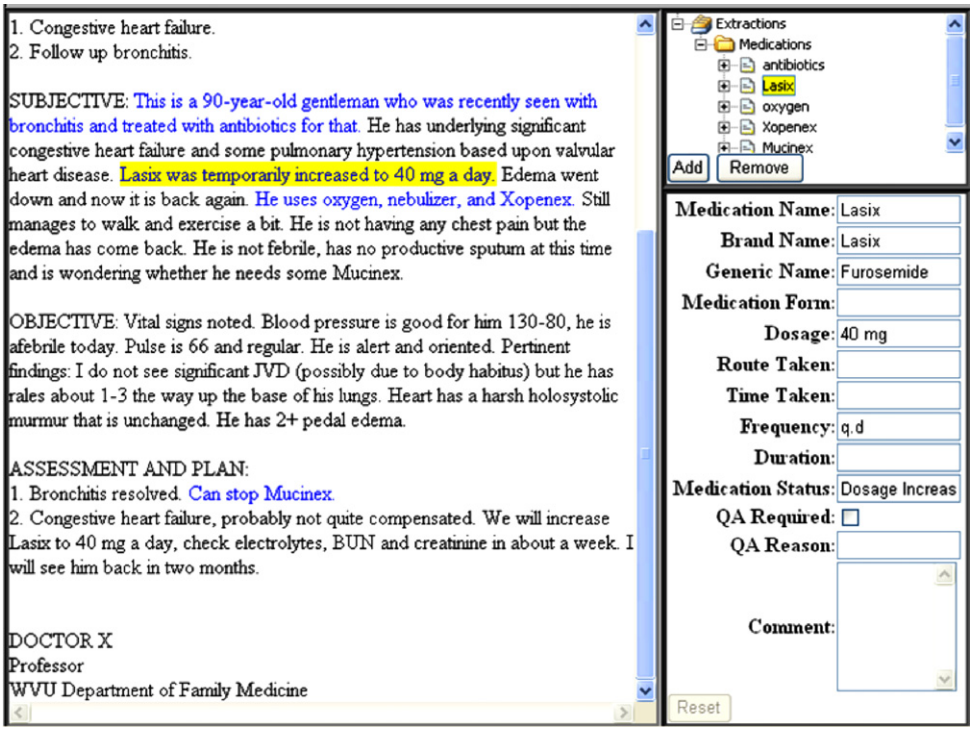

Fig. 1 – Medication extraction verification interface.

Summary points

What was already known before this study

- Harmful medication errors have stimulated an interest in automating medication lists and the reconciliation of medications between care environments.
- Using a smaller corpus of documents, Turchin et al. previously evaluated four NLP engines for their ability to extract medication information [18].

What this study has added to our knowledge

- We evaluated four commercial NLP engines on a large corpus of 2000 documents.
- We describe a novel methodology for creation of an automated gold standard using result sets from the four vendors.
- Commercial NLP engines cannot presently populate medication lists reliably alone, but provide a good foundation for human abstraction of medication information.

of our second evaluation comparing individual vendor results versus the automated comparison standard. This latter analysis could have further highlighted similarities and differences between the four vendor NLP products.

5. Conclusion

Although automated extraction may serve as the foundation for a manual review process, it is not ready to automate medication lists without human intervention. This observation is similar to the experience of vendors that use NLP to assist in medical coding. The field of computer assisted coding uses NLP to assign codes such as ICD9 and CPT and a manual review process verifies the accuracy of the assignment [23]. NLP promises to facilitate the extraction of medication information from dictated physician notes but at the current time, it requires manual verification of the results.

Acknowledgements

We were happy to have received the co-operation of WVU Hospitals and in particular: Melisa Martin and Patricia Wilson of Health Information Management, and Mark Combs, Mike Denney, and Jeff Cox of Information Technology. And, we also would like to thank Mark Ivie of MedQuist for sponsoring this effort. The prototype extraction interface shown in Fig. 1 was developed by a team of MedQuist personnel—Chen He, Tad Davis, Linda Crossley, Timothy Owensby and David Taylor. Chris Richards helped in de-identifying the records.

We also would like to thank the following vendors who contributed their results of extraction for this effort: Language and Computing Inc.; Coderyte, LingoLogics and Artificial Medical Intelligence.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ijmedinf.2008.08.006.

REFERENCES

[1] N. Sager, C. Friedman, M.S. Lyman, Medical Language Processing: Computer Management of Narrative Data, Addison-Wesley, 1987.
[2] C. Friedman, Towards a comprehensive medical language processing system: methods and issues, in: Proceedings of AMIA, Annual Fall Symposium, 1997, pp. 595–599.
[3] G. Hripcsak, G.J. Kuperman, C. Friedman, Extracting findings from narrative reports: software transferability and sources of physician disagreement, Meth. Inform. Med. 37 (January(1)) (1998) 1–7.
[4] S. Meystre, P.J. Haug, Automation of a problem list using natural language processing, BMC Med. Inform. Decis. Mak. 5 (August) (2005) 30.
[5] B. Hazlehurst, H.R. Frost, D.F. Sittig, V.J. Stevens, MediClass: a system for detecting and classifying encounter-based clinical events in any electronic medical record, J. Am. Med. Inform. Assoc. 12 (September–October(5)) (2005) 517–529.
[6] B.W. Mamlin, D.T. Heinze, C.J. McDonald, Automated extraction and normalization of findings from cancer-related free-text radiology reports, in: AMIA Annual Symposium Proceedings, 2003, pp. 420–424.
[7] D.T. Heinze, M.L. Morsch, J. Holbrook, Mining free-text medical records, in: Proceedings of AMIA Symposium, 2001, pp. 254–258.
[8] K. Denecke, J. Bernauer, Extracting specific medical data using semantic structures. Lecture Notes in Computer Science, vol. 4594/2007, Artificial Intelligence in Medicine, Springer, Berlin/Heidelberg, 2007, pp. 257–264.
[9] C. Friedman, L. Shagina, Y. Lussier, G. Hripcsak, Automated encoding of clinical documents based on natural language processing, J. Am. Med. Inform. Assoc. 11 (September–October(5)) (2004) 392–402.
[10] D. Heinze, M. Morsch, et al. LifeCode®—A Deployed Application for Automated Medical Coding, AI Magazine, Summer, 2001.
[11] W. Morris, D. Heinze, et al., Assessing the accuracy of an automated coding system in emergency medicine, in: Proceedings of the 2000 AMIA Annual Fall Symposium, November, 2000.
[12] P. Resnik, M. Niv, et al., Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding, in: Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings, Fall, 2006.
[13] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (September–October(5)) (2007) 550–563.
[14] C. Clark, K. Good, et al., Identifying smokers with a medical extraction system, J. Med. Inform. Assoc. 15 (January–Februaty(1)) (2008) 26–39.
[15] JCAHO, 2008 National Patient Safety Goals, JCAHO Website available at: http://www.jointcommission.org/PatientSafety/NationalPatientSafetyGoals/08_cah_npsgs.htm (accessed February 4, 2008).
[16] E.G. Poon, B. Blumenfeld, et al., Design and implementation of an application and associated services to support interdisciplinary medication reconciliation efforts at an

integrated healthcare delivery network, J. Am. Med. Inform. Assoc. 13 (November–December(6)) (2006) 581–592.
[17] J.J. Cimino, T.J. Bright, J. Li, Medication reconciliation using natural language processing and controlled terminologies, in: K. Kuhn, et al. (Eds.), MEDINFO, 2007, pp. 679–683.
[18] A. Turchin, L. Morin, et al., Comparative evaluation of accuracy of extraction of medication information from narrative physician notes by commercial and academic natural language processing software packages, in: Proceedings of AMIA, 2006, pp. 789–793.
[19] D. Gupta, M. Saul, J. Gilbertson, Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research, Am. J. Clin. Pathol. 121 (February(2)) (2004) 176–186.
[20] W.W. Chapman, J.N. Dowling, D.L. Chu, ConText: an algorithm for identifying contextual features from clinical text, in: BioNLP Workshop of the Association for Computational Linguistics Prague, Czech Republic, 2007, pp. 81–88.
[21] W.W. Chapman, W. Bridewell, et al., Evaluation of negation phrases in narrative clinical reports, in: Proceedings of AMIA Symposium, 2001, pp. 105–109.
[22] R.H. Dolin, L. Alschuler, et al., The HL7 clinical document architecture, release 2, J. Am. Med. Inform. Assoc. 13 (1) (2006) 30–38.
[23] Y. Jiang, M. Nossal, P. Resnik, How does the system know it's right? Automated confidence assessment for compliant coding, in: The CAC Proceedings, Perspectives in Health Information Management, Fall, 2006, http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_032079.html.