Comparing Concept Normalization Accuracy and Speed for Medical Problems and Medication Allergies

Stephane Meystre, MD, PhD1,2, Andrew Trice, BS2, Youngjun Kim1, PhD, Paul Heider, PhD1
1Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC
2Clinacuity, Inc., Charleston, SC

Introduction: Electronic Health Record (EHR) systems growth results in very large quantities of patient data becoming available in electronic format. Secondary use of this patient data is essential to fulfill the potentials for effective clinical research, high quality healthcare, and improved healthcare management. Instead of relying on existing but often biased and insufficiently detailed codes assigned for reimbursement and administrative purposes, high quality and efficient healthcare and effective clinical research need accurate and detailed clinical data that can only be found in patient EHRs. Trials based on this clinical ‘big data’ are even expected to reshape clinical research.1 For flexibility, expressiveness, efficiency, and historical reasons, most detailed clinical data found in EHRs is captured in free-text, with neither structure nor coding.

To extract information from this unstructured text, a dictionary lookup task, which links text to standard terminologies, is typically required. This task consists in searching the text for mentions of concept terms from the ‘dictionary’ (i.e., standard terminology). It is sometimes also called “concept normalization” or even “concept recognition” even if the latter would be closer to “named-entity recognition” or “entity recognition” (i.e., simply finding mentions of some categories of entities without linking them to specific concepts in some terminology, like finding that “diabetes mellitus” is a Disease, but not that it maps to [C0011849; Diabetes Mellitus] in the UMLS Metathesaurus). Most natural language processing (NLP) software applications used with EHR text include some dictionary lookup. These applications include prominent examples such as MetaMap,3 MedLEE4 (now commercially available as REVEAL, from Health Fidelity, Palo Alto, CA), NOBLE Coder,5 the NCBO Annotator,6 Textractor,7 or Apache cTAKES.8,9 Some use their own dictionary lookup algorithm while others use existing algorithms like Apache Lucene10 or Apache UIMA ConceptMapper.11

In the context of a larger project to automatically extract information from the EHR with high accuracy to then improve the completeness and timeliness of lists of medical problems and allergies, we examined the opportunity to improve or replace a dictionary lookup module based on Apache Lucene with a focus on accuracy and speed.

Methods: A variety of dictionary lookup tools, text corpora, and dictionaries were used in this study, as explained below. Dictionary lookup tools: We compared Apache Lucene, ConceptMapper, and cTAKES (fast lookup). Lucene is a popular and powerful text search engine library used by numerous websites and applications (e.g., LinkedIn and Twitter). It is used in Textractor7 and in our prototype application for extracting medical problems and allergens from clinical notes.12 In that application, Lucene is combined with a normalization process that includes abbreviation expansion, stemming, removal of punctuation, lowercasing, reordering of tokens and removal of stopwords. On top of word tokens, it is fed noun phrase chunks and named entities detected by a machine learning classifier. To allow for even comparisons between tools, we also evaluated a simpler implementation of Lucene without this normalization process. ConceptMapper is a dictionary lookup tool distributed as part of Apache UIMA's suite of tools. It is also a powerful and highly configurable tool, capable of non-contiguos terms mapping and fast performance.

Apache cTAKES is a popular open source clinical NLP application built on Apache UIMA. It offers a fast dictionary lookup module in its latest version (4.0).

Clinical text corpora: Two collections of clinical text notes were used in this study. Both contain a variety of notes randomly selected from patient populations at the University of Utah (Salt Lake City, UT) and at the Medical University of South Carolina (MUSC, Charleston, SC). The Utah collection includes 770 notes and the MUSC collection 522 notes. Both collections were automatically de-identified and manually annotated for a selection of medical problems and allergens.12

Dictionaries: Two different dictionaries were used. The first is the Clinical Observations Recordings and Encoding (CORE) problem list subset of SNOMED-CT, a collection of concepts used for coding problem list entries or other summary documentation.13 It includes 6,117 concepts with 106,616 terms. The second was semi-automatically built to focus on the selection of medical problems and medications causing allergies used for annotating both corpora. A seed list of manually-curated concepts was automatically augmented with narrower/child concepts and other relations found in the UMLS Metathesaurus. This ‘custom’ dictionary includes 24,833 concepts with 134,408 terms.

Results: Speed metrics included seconds per note and seconds per 5,000 characters to account for note size differences between corpora. As seen in Table 1, ConceptMapper was the fastest, about 1.5-7 times faster than cTAKES fast lookup and about 160-780 times faster than Lucene with normalization. Without normalization, Lucene was faster with times between ConceptMapper and cTAKES fast lookup.

Accuracy metrics included recall, precision, and the F1-measure. Accuracy was assessed in two separate steps: 1) concept mentions in text identification and 2) identified mentions normalization (i.e., mapping to standard terminologies) accuracy.
Concept mentions identification: When only considering the identification of mentions of problems or allergens with overlapping text spans, recall ranged from 59.65% (with cTAKES, MUSC corpus, and SNOMED CORE) to 91.52% (with ConceptMapper, Utah corpus, and SNOMED CORE), as seen in Figure 1. By design, our dictionary lookup was identifying all mentions of medications rather than only the ones causing allergy. A subsequent classifier not included in this study then focuses only on medications causing allergy. Precision measurements were therefore low, between 8.41% and 86.03%.

Identified concepts normalization: When assessing the mapping of mentions of problems or allergens identified in the previous step with UMLS concepts, the F1-measure ranged from 84.2% to 97.49% (Figure 2).

All evaluations were realized on Apple MacBook Pro (Core i7 2.8 GHz processors, 16 GB of RAM) laptop computers.

**Acknowledgments**: Research supported by the National Cancer Institute (R41CA180190).

**References**


---

**Table 1: Dictionary lookup tools speed**

<table>
<thead>
<tr>
<th>Tool</th>
<th>UT Utah corpus</th>
<th>MUSC Utah corpus</th>
<th>UT MUSC corpus</th>
<th>MUSC MUSC corpus</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucene v7.7 normalized</td>
<td>0.924</td>
<td>0.797</td>
<td>2.264</td>
<td>2.356</td>
<td>1.647</td>
</tr>
<tr>
<td>Lucene v7.7 no normalization</td>
<td>0.008</td>
<td>0.007</td>
<td>0.009</td>
<td>0.009</td>
<td>0.008</td>
</tr>
<tr>
<td>ConceptMapper</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>cTAKES fast lookup</td>
<td>0.022</td>
<td>0.019</td>
<td>0.023</td>
<td>0.023</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Average size of notes in each corpus: Utah 5,793 characters, MUSC 4,803 characters.