Generating summaries of hospitalizations: A new metric to assess the complexity of medical terms and their definitions

Sabita Acharya, Barbara Di Eugenio, Andrew D. Boyd, Karen Dunn Lopez, Richard Cameron, Gail M Keenan
University of Illinois at Chicago
Chicago, IL, USA

Abstract

Our system generates summaries of hospital stays by combining information from two heterogenous sources: physician discharge notes and nursing plans of care. It extracts medical concepts from both sources; concepts that are identified as "complex" by our metric are explained by providing definitions obtained from three external knowledge sources. Finally, relevant concepts (with or without definition) are realized by SimpleNLG.

1 Introduction

In the US, about 42 million people are hospitalized every year (Adams et al., 2013). When patients are released, they often do not understand their discharge instructions and what happened to them in the hospital (Haatainen et al., 2014). Our solution is to generate a concise summary that integrates the separate physician and nursing documentations, since in current hospital practice, no comprehensive record exists of the care provided to a patient.

After summarizing our baseline work previously reported in (Di Eugenio et al., 2014), we focus on medical term complexity. The novelty of our work consists in the multiprong approach underlying our complexity metric, that includes linear regression and clustering; and in applying the metric not just to the term in question, but to its many available definitions, so as to choose the simplest one to refer the patient to.

2 Related Work

Only few NLG systems generate personalized information from medical data for the *patient*, as opposed to health care personnel (Williams et al., 2007; Mahamood and Reiter, 2011). As concerns identifying difficult terms, some applications search for them in vocabularies or in specific corpora (Ong et al., 2007; Kandula et al., 2010). The drawback of these approaches is that they make an underlying assumption that all the terms that appear in such resources are *complex* and need to be explained further. Moreover, since none of the currently available vocabularies/corpora are exhaustive enough, this method is not reliable. Our approach for identifying complex terms is closer to (Shardlow, 2013), but we are interested in medical terms and use five times as many features, and a two-step approach, not their single SVM model. Similar to (Ramesh et al., 2013), we provide definitions for terms; but Ramesh et al. consider every term whose semantic type falls within a set of 16 types derived from the Unified Medical Language System[1] (UMLS) as complex, while we don't make such assumptions.

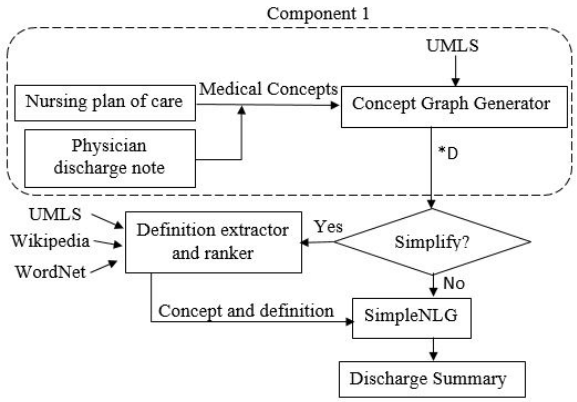
3 System Workflow

In our previous work (Di Eugenio et al., 2014), we set up the core of the NLG pipeline represented by *Component 1* in Figure 1. We also computationally demonstrated that doctor and nurses focus on different aspects of care (Di Eugenio et al., 2013; Roussi et al., 2015), and hence, that both perspectives need to be included. The first input to the system is the *hospital course* section of the doctor's free text discharge notes.[2] Medical concepts are extracted from the discharge notes by MedLEE (Friedman et al., 2004), a medical information extraction tool that maps entities to concepts in UMLS. UMLS includes 2.6 million concepts, identified by Concept Unique Identifiers (CUIs). A concept is described by either

[1] http://www.ncbi.nlm.nih.gov/books/NBK9676/

[2] The de-identified notes come from our hospital.

Proceedings of The 9th International Natural Language Generation conference, pages 26–30,
Edinburgh, UK, September 5-8 2016.



*D: NANDA, NIC, NOC, doctor node concepts at distance 1 or 2, or intermediate nodes that connect to doctor's node

Figure 1: Schematic representation of the system

a single word or multiple words; eg., *Cerebrovascular accident* is a concept with CUI *C0038454*.

The second input to our system is structured nursing documentation as recorded via the HANDS tool (Keenan et al., 2002). HANDS employs structured nursing taxonomies (NNN, 2014):NANDA-I for nursing diagnoses, NOC for outcomes and NIC for interventions. HANDS also uses a scale from 1 to 5 to indicate the initial state of the patient for that outcome when s/he was admitted, and the expected rating at discharge. Since the nursing terminologies are already included within UMLS, they also have corresponding CUIs.

To generate the summary, for each patient, we build a graph, starting from two sets of CUIs: those extracted from the discharge notes; and those corresponding to the NANDA-I, NIC and NOC terms from HANDS. We grow the graph by querying UMLS for CUIs that are related to each of the CUIs in the initial sets. From the graph, we select those CUIs that either belong to one of the source lists, or are required to form a connection between a doctor-originated concept and a nurse-originated concept that would otherwise remain unconnected. In Figure 2, *difficulty walking* is a NANDA-I diagnosis that is related to *nervous system disorder*, which is an intermediate node discovered by our graph building procedure. Concepts corresponding to the selected CUIs are candidates for inclusion in our summary. First, a filter identifies whether the concept is *Simple* or *Complex*. If it is identified as *Complex*, it is sent to the *Definition extractor and ranker* module that retrieves definitions of the concept from three external knowledge sources (see Section 5), ranks them according to their increasing complexity, and returns the simplest definition. These concepts, along with relevant verbs (that are supplied depending upon whether the concept is a diagnosis/ intervention/ treatment/ intermediate node) are couched as features of phrasal constituents via the operations provided by the SimpleNLG API (Gatt and Reiter, 2009), which then assembles grammatical phrases in the right order.

Our running example summary is shown in Figure 3. So far, we have generated discharge summaries for 58 patient cases; the average number of concepts in a summary is 33. Out of all the concepts that appear in our 58 summaries, 20% consist of a single word, 52% of two words, 16% of three words, and 12% of more than 3 words. Instead of explaining each word in a concept, we provide a definition for the concept as a whole. In the following, we will more specifically refer to concepts as "terms".

4 Term complexity assessment

Most of the earlier work assumes that every medical term is complex, and maps it to a simpler term via lexica (Ong et al., 2007). First, it is too simplistic to assume that every medical term is complex, however no measure exists to assess the complexity of a medical term. Tools for assessing health literacy (REALM, TOFHLA, NAALS) and reading level (Flesch, Fry Graph, SMOG) work only on sentences and not on words (CHIRR, 2012).

Second, as concerns the coverage of existing vocabularies for replacing complex terms, we started by assessing the foremost resource currently available, the Consumer Health Vocabulary (CHV) (Doing-Harris and Zeng-Treitler, 2011), which maps medical terms to plain language expressions. We found out that CHV provides a simplified alternative for only 14% of our terms, most of which we contend are not "simple" enough. We also compiled several vocabulary sources found online: MedicineNet[3], eMedicine[4], MedlinePlus[5] into a single lexicon, but only 2.17% of the medical terms

[3] www.medicinenet.com/ medterms-medical-dictionary

[4] www.emedicinehealth.com/medical-dictionary-definitions

[5] www.nlm.nih.gov/medlineplus

You were admitted for acute subcortical cerebrovascular accident. Difficulty walking related to nervous system disorder was treated with body mechanics promotion. Mobility as a finding has improved significantly and outcome has met the expectation. Risk for Ineffective Cerebral Tissue Perfusion was treated with medication management and administration:oral.[...] As a result, risk control behavior: cardiovascular health has improved slightly. Verbal impairment related to communication impairment was treated with speech therapy. [...] As a result, fall prevention behavior and knowledge level: fall prevention have improved slightly. Disease Process, Medication, and Disease Process (Heart disease) were taught.

Figure 2: Part of version 1 of the summary for Patient 149

You were admitted for acute subcortical cerebrovascular accident. During your hospitalization, you were monitored for chances of ineffective cerebral tissue perfusion, risk for falls, problem in verbal communication and walking. We treated difficulty walking related to nervous system disorder with body mechanics promotion. Mobility as a finding has improved appreciably. We provided treatment for risk for ineffective cerebral tissue perfusion with medication management and medication administration. As a result, risk related to cardiovascular health has reduced slightly. We worked to improve verbal impairment related to communication impairment with speech therapy. As a result, communication has improved slightly. We treated risk for falls by managing environment to provide safety. We provided information about fall prevention. As a result, fall prevention behavior and fall prevention knowledge have improved slightly. With your nurse and doctors, you learned about disease process and medication.

Figure 3: Version 2 of the summary for Patient 149

from our summaries were present in them.

4.1 Measuring term complexity

In order to develop a metric for determining the complexity of terms, we need a training set of *Simple* and *Complex* terms. For this purpose: 1) We randomly selected 300 terms from the *Dale-Chall List*, which consists of 3,000 terms that are known to be understood by more than 80% of 4th grade students (DC , 2016) and labeled them as *Simple*. 2) We randomly selected 300 medical terms present in our database of 3164 terms explored by the *Concept Graph Generator* in Figure 1 for 58 patients. Two non-native undergraduate students who have never had any medical conditions were asked to annotate the 300 terms taken from our database as *Simple* or *Complex* (Cohen's Kappa k=0.786). Disagreements between the annotators were resolved via mutual consultation.

Several features were extracted for each of the 600 terms: a) Lexical features: number of vowels, consonants, prefixes, suffixes, letters, syllables per word. b) Count of each type of POS, i.e. number of nouns, verbs, adjectives, prepositions, conjunctions, determiners, adverbs, numerals (extracted by the Stanford parser) c) whether the term is present in Wordnet d) UMLS derived features: number of semantic types, synonyms, and CUIs that are identified for the term; whether the term is present in CHV; whether the entire term has a CUI; whether the semantic type of the term is one of the 16 semantic types from (Ramesh et al., 2013).

As a first step, linear regression was performed on the 600 terms with *Complexity* (0-Simple, 1-Complex) as the dependent variable. This process filtered out unimportant features for predicting complexity: number of letters, consonants; number of prepositions, conjunctions; 4 out of the 16 semantic types discussed above: *Cell or Molecular Dysfunction*, *Experimental Model of Disease*, *Finding*, and *Physiologic Function*. It also provided a linear regression function that hence includes only the important features, which we will collectively call **F**.

As a second step, Expectation-Maximization clustering was performed on the remaining 2864 terms from our database, using the earlier collected 600 terms as *cluster seeds*. This resulted in 3 clusters. Of the 600 cluster seeds, 70% of those in Cluster1 had *Simple* label; 79% of those in Cluster3 had *Complex* label; 58% of those in Cluster2 had *Simple* label and 42% had *Complex* label. This indicates the presence of three categories of terms: some that can be identified as *Simple* (Cluster1), some that are *Complex* (Cluster3), and the rest for which there is no clear distinction between *Simple* and *Complex* (Cluster2). For the terms in each of these clusters,

we further supplied feature values from the set **F** to the linear regression function and analyzed the corresponding scores. We found out that across all clusters, 88% of the terms labeled as *Simple* have scores below 0.4 while 96% of the terms whose score was above 0.7 were labeled *Complex*. For the terms whose score was between 0.4 and 0.7, no clear majority of *Simple* or *Complex* labeled terms was observed in any of the clusters. This further verifies the observation made during clustering that our dataset consists of three categories of terms. The thresholds of 0.4 and 0.7 were obtained by sorting the scores of the terms within each cluster and looking for the highest difference in consecutive scores.

Hence, given a new term to assess, our system will: a) Extract features **F** b) Supply feature values to the linear regression function c) If the score is below 0.4, the term is considered *Simple*; if score is above 0.7, the term is considered *Complex* and a definition is provided. For scores between 0.4 and 0.7, definition will be provided only if the term's semantic type falls within our list of 47 semantic types, obtained after removing non-medical types like *Organization* from the list of 133 semantic types in UMLS.

5 Choosing an appropriate definition

For the terms that are identified as *Complex* by our metric, we will extract definitions from three external knowledge sources: Wikipedia (extract only the first sentence), WordNet, and UMLS. Since more than 60 vocabulary sources are integrated into UMLS, a single term might have multiple definitions. Hence, definitions from all the three sources are obtained and for each definition, medical concepts present in it are extracted. Using our metric for determining complexity (Section 4.1), we obtain scores for each of the concepts in a definition and add them together to get a single score. The definition with the lowest score is eventually chosen.

For instance, for a term *Cerebrovascular accident*, 1) our metric returns a score of *0.801*, which indicates that a definition needs to be supplied. 2) *Definition extractor and ranker* module extracts definitions of the term from three knowledge sources and ranks them. 3) The definition from Wikipedia has the lowest score and hence the first occurrence of the term *Cerebrovascular accident* in our summary will have the definition *when poor blood flow to the brain results in cell death* attached to it. All the terms that have been highlighted in Figure 3 are found to be *Complex* and a corresponding definition is provided by the system. These definitions can be presented in different forms (like *footnote* or *tooltip text*) depending upon the medium in which the summary is going to be presented. Whereas we have not run a formal evaluation, two of our patient advisors observed that our current summaries have vastly improved compared to the baseline.

6 Current and Future Work

Currently, some of the terms like *central venous* and *organism strain* are identified as *Simple* by our metric. In order to improve the accuracy of our metric, we plan to add a feature that represents the frequency of a term in Google-ngram corpus as is done in (Grabar et al., 2014; Kauchak and Leroy, 2016) and evaluate its effectiveness in predicting complexity. This could also be useful in disambiguating the complexity of terms with score between 0.4-0.7.

Our next immediate goal is to include the patient's perspective in our summaries, similarly to Gkatzia et al. (2014). We are collecting open-ended interviews with 40 patients and have interviewed four so far. We are currently transcribing the recordings; we will code them for features of interests, and plan to mine them with methods appropriate for *small data* (Smith et al., 2014). Once summaries can be personalized, we plan to perform first, controlled evaluations, and eventually longer-term assessments of whether our summaries engender better health, i.e., by better adherence to medications.

References

P.F. Adams, W.K. Kirzinger, and Martinez M. 2013. *Summary Health Statistics for the U.S. Population: National Health Interview Survey, 2012*, volume 10 of *Vital and Health Statistics*. Centers for Disease Control and Prevention.

CHIRR. 2012. Health literacy. Consumer health informatics research resource, https://chirr.nlm.nih.gov/health-literacy.php.

2016. Readability Formulas. readabilityformulas.com.

Barbara Di Eugenio, Camillo Lugaresi, Gail M. Keenan, Yves A. Lussier, Jianrong Li, Mike Burton, Carol

Friedman, and Andrew D. Boyd. 2013. HospSum: Integrating physician discharge notes with coded nursing care data to generate patient-centric summaries. In *AMIA 2013, American Medical Informatics Association Annual Symposium*, Washington D.C., November. Abstract.

Barbara Di Eugenio, Andrew D. Boyd, Camillo Lugaresi, Abhinaya Balasubramanian, Gail Keenan, Mike Burton, Tamara Goncalves Rezende Macieira, Jianrong Li, Yves Lussier, and Carol Friedman. 2014. PatientNarr: Towards generating patient-centric summaries of hospital stays. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 6–10, Philadelphia, Pennsylvania, U.S.A., June. Association for Computational Linguistics.

K.M. Doing-Harris and Q. Zeng-Treitler. 2011. Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of Medical Internet Research*, 13(2).

C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. 2004. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392.

A. Gatt and E. Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.

D. Gkatzia, V. Rieser, A. McSporran, A.R. McGowan, A.R. Mort, and M. Dewar. 2014. Generating verbal descriptions from medical sensor data: A corpus study on user preferences. *BCS Health Informatics Scotland. Glasgow, UK*.

Natalia Grabar, Thierry Hamon, and Dany Amiot. 2014. Automatic diagnosis of understanding of medical words. *EACL 2014*, pages 11–20.

K. M. Haatainen, Ta. Tervo-Heikkinen, and K. Saranto. 2014. Adult patients' experiences of discharge education in an emergency department: a systematic review protocol. *The JBI Database of Systematic Reviews and Implementation Reports*, 12(5):80–87.

S. Kandula, D.y Curtis, and Q. Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *AMIA Annu Symp Proc*, volume 2010, pages 366–70.

David Kauchak and Gondy Leroy. 2016. Moving beyond readability metrics for health-related text simplification. *IT Professional*, 18(3):45–51.

G.M. Keenan, J.R. Stocker, A.T. Geo-Thomas, N.R. Soparkar, V.H. Barkauskas, and J.A.N.L. Lee. 2002. The HANDS Project: Studying and Refining the Automated Collection of a Cross-setting Clinical Data set. *CIN: Computers, Informatics, Nursing*, 20(3):89–100.

S. Mahamood and E. Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21, Nancy, France, September. Association for Computational Linguistics.

2014. NNN: Knowledge-based terminologies defining nursing. http://www.nanda.org/nanda-i-nic-noc.html.

E.l Ong, J. Damay, G. Lojico, K. Lu, and D. Tarantan. 2007. Simplifying text in medical literature. *Journal of Research in Science, Computing and Engineering*, 4(1):37–47.

Balaji Polepalli Ramesh, Thomas K Houston, Cynthia Brandt, Hua Fang, and Hong Yu. 2013. Improving patients' electronic health record comprehension with noteaid. In *MedInfo*, pages 714–718.

Khawllah Roussi, Vanessa Soussa, Karen V Dunn Lopez, Abhinaya Balasubramanian, Gail M Keenan, Michel Burton, Neil Bahroos, Barbara Di Eugenio, and Andrew Boyd. 2015. Are we talking about the same patient? In *IOS Press*.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *ACL (Student Research Workshop)*, pages 103–109. Citeseer.

G. CS Smith, S. R Seaman, A. M Wood, P. Royston, and I. R White. 2014. Correcting for optimistic prediction in small data sets. *American Journal of Epidemiology*.

S. Williams, Pa.l Piwek, and R. Power. 2007. Generating monologue and dialogue to present personalised medical information to patients. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 167–170, Saarbrücken, Germany, June.