Knowledge Node and Relation Detection

Jain Qin[1][0000-0002-7094-2867], Bei Yu[1][0000-0001-5425-0011] and Liya Wang[1]

[1] Syracuse University, Syracuse, NY, USA
{jqin, byu, lwang32}@syr.edu

Keywords: Knowledge representation; Knowledge node relations; Knowledge node detection; Relation recognition; MetaMap; SemRep.

Abstract. A bottleneck problem in detecting knowledge nodes and their relations is how to extract accurately and correctly and codify the complex knowledge assertions from full-text documents (human intelligence) into the format of "machine intelligence" (computer-processable knowledge assertions). This paper reports a preliminary study that aims at this bottleneck problem by starting from the fundamentals of KR—representing knowledge from full-text documents by using knowledge node and relation recognition methods and tools. We collected data from full-text biomedical research publications and used manual and automatic tools to investigate the strengths and limitations of these methods. The findings show that MetaMap did a better job in detecting concepts from texts while SemRep is capable of extract relations between k-nodes. The paper presents the findings from the perspectives of degree of abstraction, types of k-nodes and relations, and linguistic structures and the evaluation results using the BLEU and cosine similarity measures.

1 Introduction

Document content is typically represented by keywords and/or terms from knowledge organization systems (KOS). Although numerous algorithms and models, for example, the vector space model (Salton et al., 1975) and the Latent Semantic Analysis (LSA) model (Deerwester et al., 1990), have been developed for document representation over the last 40 years, the term-based representation of document content is facing new challenges from the changing research landscape in which data and documents are deeply interdependent and interconnected. It has become common practice in data repositories to link datasets with associated publications (e.g., GenBank and Dryad). Much of these challenges is due to the inherent limitations of term-based document representation, that is, topics, concepts, and entity names are represented by discrete terms, and the relations between these terms are not present at the time of indexing, but rather, rely on the relations established in KOS, e.g., concept scope terms such as broader, narrower, or associated terms, or linguistic relations such as synonymous terms or qualifiers for concepts. The meager relation types in KOS and lack of rich semantic relations beyond concept scope and linguistic relations among terms not only make it difficult for human users to effectively discover data and information and comprehend the complex knowledge network, but also for information retrieval systems to fully utilize the advances in digital technologies and benefits of big data and "big knowledge" (Perl et al., 2017).

One way to address the challenges facing term-based document representation is to shift the representation approach from terms to knowledge networks consisting of nodes and relations. This means a transition from term-centric representation to concept-relation representation: terms as knowledge nodes (or k-nodes) are linked by relations to form a knowledge network. For example, influenza is a single concept representing a kind of disease, H1N1 as a kind of virus is a single concept, and the two concepts are related because H1N1 virus causes influenza, which may be written as a knowledge assertation (H1N1, causes, influenza). If we analyze it further, we can find that this assertion implies two separate assertions: (H1N1, IS-A, virus) and (influenza, IS-A, disease), hence the complete assertion may be expressed as [(H1N1, IS-A, virus), causes, (influenza, IS-A, disease)]. These assertions (or factual knowledge) can be generalized in a triple format (concept1, relation, concept2) or (subject, predicate, object), whereas concepts and relations may be complex and nested. This example is used to demonstrate a need to reexamine the nature of indexing or document content representation from an epistemological approach to enable richer, more semantic representations.

Representing knowledge assertions has its roots in artificial intelligence (AI). In the short history of AI, knowledge representation formats such as predicate logic, frame systems, semantic networks, and rule systems from the official production system family have been popular (van Hamelen et al., 2008). However, they are difficult to be generated automatically. Using triples for representing knowledge assertions is a popular method in current representation technology. Knowledge repositories (or datasets) that contain assertions have been developed, and DBpedia and Google's Knowledge Graph are well-known examples. Triple stores and ontologies have been created and in use in national libraries, including Linked Data Service at the Library of Congress (LC) and the Unified Medical Language System (UMLS) at the National Library of Medicine (NLM) and Getty Research Institute Linked Data Service. Studies are also being done to explore how the vast knowledge discovered from big data can be easily comprehended and properly and creatively used by human users (Perl et al., 2017). While shifting from a term-centric paradigm to a k-node-relation representation presents promising prospects for innovating data and document representation as well as information and data discovery, the task is highly challenging and requires the orchestration of several fields' techniques and methods, including natural language processing (NLP), knowledge representation (KR), Semantic Web (SW) technologies, and knowledge organization systems.

A bottleneck problem in this paradigmatic shift lies in detecting k-nodes and their relations accurately and correctly and codifying complex knowledge assertions from full-text documents (human intelligence) into the format of "machine intelligence" (computer-processable knowledge assertions). This paper explores the epistemological and linguistic aspects of this bottleneck problem by starting from the fundamentals of KR—representing knowledge from full-text documents by using k-node and relation recognition methods and tools.

2 Relevant literature

Detecting knowledge nodes and relations in full-text documents automatically and representing them in some degree of formality is similar in many ways to knowledge representation in AI. What is knowledge representation then? The notion of KR implies five distinct roles according to Davis et al. (1993): it is a surrogate or a substitute for the thing itself, a set of ontological commitments, a fragmentary theory of intelligent reasoning, a medium for pragmatically efficient computation, and a medium of human expression of things about the world. Davis et al. further indicate that each of these roles is "important to the nature of representation and its basic tasks" and "all five aspects are essential representation issues" (Davis et al., 1993, pp. 31-32). By adopting a KR approach in representing document content, we argue that knowledge node and relation detection by nature is a process of using terminology (surrogate) to express the assertion knowledge in text, which is committed to be focused on the part of knowledge as objectively and accurately as our capability allows, and such terminology will be in some formality to facilitate efficient computation and human expression of knowledge latent in the text.

Research in KR in the information science field appears to have followed two different but intertwined approaches, both having a great deal to do with the data sources under study. The first approach uses KOS as the data source to transform the concepts defined in KOS into knowledge maps or graphs. In order to make such transformation possible, a key task is to establish relations between concepts and the existing concept scope and linguistic relations become handy in defining the transformation data models, as seen in the Simple Knowledge Organization System (SKOS) (Mills & Brickley, 2005). Traditional thesauri such as Medical Subject Headings (MeSH) and Library of Congress Subject Headings (LCSH) were converted into the Resource Description Framework (RDF) format (NLM, 2018; LC, 2018). Because the concepts and relations are already defined in these KOS, such transformation's main task is to remodel the KOS as ontologies. In MeSH's case, remodeling MeSH into RDF format means that they must address questions such as: How should MeSH RDF classes and hierarchical relations be expressed? How should it handle Descriptor-Qualifier combinations? The project team designed a customized RDF data model to address these questions (Bushman et al., 2015). The resulting MeSH RDF Data Model is a good illustration of the five roles of KR.

Knowledge representation also occurs in transforming legacy data models (e.g., relational database, XML) into RDF or Web Ontology Language (OWL) format or integrating multiple linked data repositories to build a specialized KR base. In reviewing 14 articles that cited any of the three significant linked data projects—Bio2RDF, Open PHACTS, and/or EBI RDF, Barros and Couto (2016) conclude that RDF technologies have "a strong impact on how the Life and Health Sciences community is storing, integrating and sharing data and knowledge" (p. 182). Although automatic methods such as text mining have been used to support semantic annotations to common ontologies, human intervention is still required in most cases (Barros & Couto, 2016,).

The goal of transformation of existing thesauri or controlled vocabularies into RDF encoded format is to codify the concepts and relations in more formal and better structured format for efficient computation. This approach, however, does not connect the knowledge existing in documents and research data to KOS. In other words, the fact that controlled vocabularies are transformed into RDF format only builds a semantic infrastructure to support the representation of knowledge existing in documents and data. The work of connecting the knowledge in documents and data to KOS is what we usually refer to as indexing, text categorization, document classification, or document representation. In this paper, we take document content representation a step further to formalize such representation in the form of (subject, predicate, object), that is, representing not only knowledge nodes but also the relations among the nodes.

Representing knowledge in documents has been studied extensively and generated a vast body of publications. Popular models include the vector space model (Salton et al., 1975) and text classification using machine learning algorithms such as the K nearest neighbor (Altman, 1992). The simplest model of document representation is N-gram where words are represented as strings of N length. Other approaches of document representation include single word representation, stemmed single word representation, phrases, and rich document representation (Keikha et al., 2008). No matter which model or approach is used, documents must be processed with NLP tools to conduct morpho-syntactic analysis, disambiguate specialty terms and general language words, normalize synonyms and entity names, and tag part of speech (POS) before feature extraction and classification can be performed (Dobrokhotov et al., 2003).

Recognizing medical concepts and their semantic relations is a fundamental task in biomedical informatics. Concept and relation recognition can support a range of tasks such as concept-based text retrieval (Zhong and Huang, 2006), literature-based discovery (Hristovski et al., 2006), etc. Prior research in biomedical national language processing (bioNLP) has resulted in a suite of concept recognition tools, utilizing knowledge-based approach or machine learning approach, or the combination of both (Shah et al., 2009). To date the widely used, general-purpose concept recognizers use knowledge-based approach, as MetaMap (Aronson et al., 2010) and NOBLE Coder (Tseytlin et al., 2016) do. Due to the wide range of domains and genres in biomedical literature, human annotations are impractical for supporting the general-purpose, machine learning-based approach to representing knowledge in literature. However, some systems managed to combine knowledge-based and machine-learning approaches on specific tasks such as identifying drug names and gene names (Tseytlin et al., 2016).

MetaMap (Aronson et al., 2010) is a widely used tool for mapping keywords in biomedical literature to noun phrases to concepts in the UMLS. MetaMap uses the SPECIALIST Lexicon and linguistic rules to determine the best mapping (Rindflesch and Fiszman, 2003). A pilot error analysis compared MetaMap's term-to-UMLS_concept mapping against the mappings manually conducted by domain experts and discovered that missing inferential or world knowledge accounted for 30% of the errors and more nuanced NLP analysis was required for improving the performance (Divita et al., 2004).

Built on the concept recognizers, another suite of tools was developed for identifying semantic relations between concepts. SemRep (Rindflesch and Fiszman, 2003; Ahlers et al., 2007; Kilicoglu et al., 2010) is an example of relation extraction tool. It defines and extracts relations such as AFFECTS, AUGMENTS, COEXISTS_WITH, DIAGNOSES, and PREDISPOSES that are defined in Semantic Network of the Metathesaurus. SemRep has been used for extracting relations in various types of biomedical documents, such as coronary catheterization reports and clinical notes (Liu et al., 2012). Although relation recognition tool such as SemRep can generate concept relations automatically from abstracts and titles, the effectiveness and quality of relation recognition is to be evaluated on wider variety of biomedical texts.

3 Methods

The main objective of this project is to gain insights into the methods and issues in knowledge node and relation recognition. Hence the data collection was not intended to be comprehensive nor representative, but rather, focused on the methodological aspect to gather information for larger scale investigation. This project continues the work reported in Qin and Zou (2017), in which the types of knowledge nodes and relations and potential applications of knowledge networks were discussed. The methods used in this project include manual coding a small number of sentences and automatic indexing of a larger collection of documents using two indexing tools.

The source documents were retrieved from PubMed by using query terms "precision medicine" in combination with break cancer, diabetes, and oncology. The selection of articles was described in Qin and Zou (2017). Among the 30 articles selected for this study, 4 articles were for breast cancer, 5 for diabetes, and 11 for oncology. Because the detection of k-nodes and relations is performed to free text, the unit of analysis is sentence for both manual and automatic methods. We randomly selected a small number (150) of sentences from an article and analyzed sentence by sentence to identify k-nodes and relations in the format [k-node(A), relation, k-node(B)]. When select k-nodes and relations, we took considerations of discriminativeness, that is, a relation needs to be distinct enough to separate it from other relations while maintaining sufficient generality so that the relation can be applied to other similar situations. We used a template table to annotate k-nodes identified from sentences. Below are two examples of the k-node and relation annotation template:

Example sentence #1: "Protein gene products that have direct roles in driving the biology and clinical behavior of cancer cells are potential targets for the development of novel therapeutics" is annotated in Table 1:

Table 1. Sample k-nodes and relations derived from example sentence #1

Node A	relation	Node B
protein gene product	drives	biology behavior of cancer cell
protein gene product	drives	clinical behavior of cancer cell
protein gene product	is-target-of	therapeutics

Example sentence #2: "Unlike most pathologic testing, which serves as an adjunct to establishing a diagnosis, the results of HER2 testing stand alone in determining which patients are likely to respond to trastuzumab, a monoclonal antibody against HER2" is annotated in Table 2:

Table 2. Sample k-nodes and relations derived from example sentence #2

Node A	relation	Node B
pathologic testing	establishes	diagnosis
HER2 testing	is-a-kind-of	pathologic testing
trastuzumab	is-a-kind-of	monoclonal antibody
HER2	responds-to	trastuzumab

The relation "have direct roles in driving" in example sentence #1 was simplified as "drives", and the pattern of k-node and relations in Table 1 can be generalized further as "factor-A drives behavior of cancer cell." In identifying k-nodes we tried to be faithful to the original concepts as much as possible, and at meantime, we also needed to derive and/or extract k-nodes that were general enough to be meaningful in as many situations as possible. This is particularly true for terms representing relations. In addition, linguistic patterns in the sentences were also noted to help generating potential NLP rules for automatic k-node detection. The manually annotated k-nodes were then matched with concepts in the Unified Medical Language System (UMLS) Metathesaurus (NLM, 2016). When there was more than one term from UMLS matching the manual k-node, we selected the best match (usually the first term in the list of results returned) and recorded it in a spreadsheet. Not all k-nodes identified from the text have a matching term in UMLS. In this case, the value for UMLS matching term was given a "none". The manual annotation resulted in 390 k-nodes in total. Searches were performed to find a matching UMLS concept for each of these k-nodes. The manual annotations were performed by two coders. Differences between the two coders were discussed and modified after agreement was reached.

Two tools were used to generate k-nodes and relations automatically. The first one is MetaMap, which was developed at NLM for recognizing UMLS concepts in text. The strength of MetaMap lies in recognizing concepts from texts and matching them with terms in UMLS. It does not perform, however, the task of detecting relations between concepts from text. SemRep was selected because of its ability to extract three-part propositions (or semantic predications) from biomedical texts (Rindflesch & Fiszman, 2003).

To obtain comparative data for the manually annotated k-nodes and relations, we ran MetaMap and SemRep on the same texts. The resulting k-nodes and relations were evaluated to find strengths and weaknesses of automatic k-node and relation detection by using the tools. These steps generated a number of datasets: 1) manually annotated k-nodes with corresponding MetaMap results, 2) k-nodes generated from MetaMap and matching UMLS concepts, 3) relations generated manually and by SemRep, and 4) evaluation scores using two algorithms: Bilingual Evaluation Understudy (BLEU) and cosine similarity for evaluating the similarity between manual and automatic k-node detection. BLEU was originally designed to evaluate a generated sentence to a reference sentence, with 1.0 as a perfect match score and 0 representing a mismatch (Papineni et al., 2002). We decided to adopt this algorithm as one of the evaluation metric because it is quick and inexpensive to calculate, easy to understand, language independent, and correlates highly with human evaluation (Brownlee, 2017). Cosine similarity measure is a widely adopted metric in document classification/clustering, information retrieval, and many other research fields to evaluate the similarity between words or string vector space ((Erk & Pado, 2008). The four datasets will address the following questions:

- To what extent manually annotated and automatically generated k-nodes and relations are similar or dissimilar?
- What are some of the patterns of agreement and/or disagreement between the two sets of results?
- How can human intelligence (human-intervened k-node and relation recognition) be translated into machine intelligence for more accurate knowledge representation?

4 Findings

Overall, the results show some clear differences in the types of k-nodes and relations between manual annotation and automatic detection. The differences are more visible when k-nodes and relations are represented in the three-part semantic predicate format. We observed three areas of differences in k-node and relation recognition: degree of abstraction or generalization, types of k-nodes and relations, and linguistic structures.

4.1 Degree of Abstraction

Semantically, the k-nodes captured by each method have a high similarity. Regardless how each concept is expressed in language, keywords for main concepts are present across three sets of results. A closer examination of resulting concepts shows that there are varying degree of abstraction or generalization between three methods. It is easy for human intelligence to determine that “monoclonal antibody against HER2” and “time from tissue removal to tissue fixation” are concepts and separating the words in any of the two phrases would cause information loss (Table 3). Among the three methods, MetaMap seems to be the better one for recognizing complex concepts involving long phrases. SemRep breaks the sentences into simple concepts with one or two words, while manual results appear to be in between the two automatic tools. In this sense, manual method and MetaMap can handle better complex concepts while SemRep generates simple k-nodes.

Table 3. Sample k-nodes generated by manual annotation, MetaMap, and SemRep

Sentence	Manually annotated k-nodes	MetaMap extracted k-nodes	SemRep extracted k-nodes
Unlike most pathologic testing, which serves as an adjunct to establishing a diagnosis, the results of HER2 testing stand alone in determining which patients are likely to respond to trastuzumab, a monoclonal antibody against HER2.	pathologic testing HER2 testing monoclonal antibody trastuzumab HER2 monoclonal antibody diagnosis	pathologic testing results of her2 testing respond to trastuzumab results of her2 testing a monoclonal antibody against her2 diagnosis	pathologic testing HER2 testing trastuzumab monoclonal antibody diagnosis
At present, several preanalytic factors, including the time from tissue removal to tissue fixation, are underappreciated as important variables that have the potential to negatively impact the consistency and reliability of HER2 testing.	time from tissue removal to tissue fixation preanalytic factor HER2 testing preanalytic factor consistency reliability	time from tissue removal tissue fixation several preanalytic factors reliability of her2 testing several preanalytic factors consistency	time removal tissue fixation factors HER2 testing consistency
Adenocarcinoma of the breast is a leading cause of cancer morbidity and mortality among women worldwide.	Adenocarcinoma of breast cancer morbidity cancer mortality	adenocarcinoma of the breast a leading cause of cancer morbidity mortality among women worldwide	Adenocarcinoma breast cancer
A major challenge faced by clinicians treating patients with breast cancer is how to best assess patient outcomes and predict the clinical course of the disease so that the most appropriate treatment regimen can be identified.	clinical course of disease patients with breast cancer patient outcomes	faced by clinicians the clinical course of the disease patients with breast cancer patient outcomes	clinicians breast cancer patient outcomes clinical course disease

Another way to look at the degree of abstraction is through the relations detected from the text. Table 4 presents the relations identified by SemRep and manual annotation. While the manual work resulted in a much larger number and variety of relations than SemRep did, SemRep has its own predefined list of relations and many of them can be found in the relation types defined by the UMLS semantic network (McCray, 2003), which is an upper-level ontology for the biomedical domain. As such, the relations from SemRep tend to be formal because of the adoption of terms from an ontology, and those from manual annotation are closer to natural language. We tried to match the manually annotated relations to those from SemRep in three categories: exact match (semantically), similar or partial match, and no match. The mapping between two sets of relation detection results It appears to have a gap between formal ontological relation detection and manual annotation. This gap is reflected in the way how a relation is constructed, for example, both prepositional phrase and verbs are used in SemRep relations while manual results contain primarily verbs.

Table 4. Relations detected by SemRep and manual annotation

Relations detected by SemRep	Relations from manual annotation		
	Exact match	Similar/Partial match	No match
AFFECTS	affects	allows, improves, impacts, promotes provides, controls	is against is essential to documents enumerates confirms assesses assays begins with demonstrates establishes harbors has identifies includes is approved by is performed by predicts responds to
IS-A	is-a	is a kind of, exists, is equivalent to, is a prototype for, is given as	
ASSOCIATED_WITH	is associated with	is-for correlates	
AUGMENTS	expands		
CAUSES	Causes, makes, determines	leads to, promotes, drives, improves	
COMPARED_WITH		is measured by, is-tested-by, measures, is-in-context	
LOCATION_OF			
METHOD_OF	is-method-for		
PART_OF	is-part-of	is a factor of, has-attribute, has condition of	
PROCESS_OF		transmits	
TREATS TREATS(INFER)	treats	mediates, has-concordance with, targets regulates	
USES	uses	is-used-for, is-used-with	

4.2 Types of K-Nodes and Relations

In the process of manual annotation, we noticed some patterns of k-node and relation structures.

1. *Simple k-node relations*: two simple k-nodes are connected by a direct relation in the form of a single verb, which may be expressed as A→B, e.g.:

   (amplification_of_HER2_gene, promotes, receptor_activation)
   (tumor, harbors, HER2_molecular_alteration)

2. *Compound k-node relations*: it is common in the text used for this study that a k-node is related to more than one k-node that may have the same or different relations. We use the expression A→(B_1…B_n) to represent a one-to-many pattern. For example, in a sentence "Therefore, over-expression of the receptor has a direct role in mediating the biologic and clinical behavior of HER2-positive tumor cells by driving their proliferation and survival", we derived:

   (overexpression_of_receptor, mediates, biology_behavior_of_HER2-positive_tumor_cells)
   (overexpression_of_receptor, mediates, clinical_behavior_HER2-positive_tumor_cells)
   (overexpression_of_receptor, drives, proliferation_of_tumor_cells)
   (overexpression_of_receptor, drives, survival_of_tumor_cells)

   Based on the expression A→(B_1, …, B_n), the above predicates can be written together if they share the same relation:

   (overexpression_of_receptor, mediates, (biology_behavior_of_HER2-positive_tumor_cells, clinical_behavior_HER2-positive_tumor_cells))

   (overexpression_of_receptor, drives, (proliferation_of_tumor_cells, survival_of_tumor_cells))

   This pattern is only applicable to the situation in which one concept is related to multiple concepts by the same relation. Once the relation changes, the k-node-relation triple also changes to become a new one.

3. *Complex k-node relations*: sometimes a sentence contains multiple k-nodes and the relations among them are not simple nor direct. More often than not, these multiple k-nodes are chained together by "bridge" k-nodes. We use the expression A→(B→C) to represent such "chain relations" where k-node A is related to C through bridge k-node B. For example, the sentence "Unlike most pathologic testing, which serves as an adjunct to establishing a diagnosis, the results of HER2 testing stand alone in determining which patients are likely to respond to trastuzumab, a monoclonal antibody against HER2" contains a complex k-node and relation and a simple k-node and relation proposition:

   (HER2_testing, determines, (patient, responds-to, trastuzumab))
   (trastuzumab, is-a, monoclonal antibody against HER2)

4.3 Linguistic Structures

While we kept relations between k-nodes in the form of verbs, the linguistic structures for candidate k-nodes appeared in a wide variety of patterns. To decide what may be meaningful and appropriate k-nodes, we needed to address two challenges. First, identifying a k-node from a long phrase (in which may contain a clause) often means to make a decision on where the cut-off point is. For instance, "Protein gene products that have direct roles in driving the biology and clinical behavior of cancer cells" is a long phrase that contains a clause modifying the k-node "protein gene product". It would not be feasible to take the whole phrase plus the clause as the k-node but cutting off the clause would cause an important information loss.

Second, as a result of the first challenge, there are situations in which candidate relation terms in non-verb or clause format need to be transformed into verbs or verb phrases to avoid the information loss (see Example 1 and Table 1 in the Methods section). Such non-verb-to-verb transformation often results in multiple k-node-relation-k-node propositions to represent the complex concepts.

In the process of manual annotation, we generalized some linguistic patterns for k-nodes:

- *Simple k-nodes*: nouns that may be mapped directly to concepts in UMLS;
- *Compound k-nodes*: noun phrases that consist of two or more nouns and each noun may find a matching concept in UMLS;
- *Complex k-nodes*: this type of k-nodes often appears in the form of noun/noun phrase + preposition + noun/noun phrase format, for example, "dysregulation of gene", "therapeutic target in clinical oncology", and "sensitivity to cytotoxic drug".

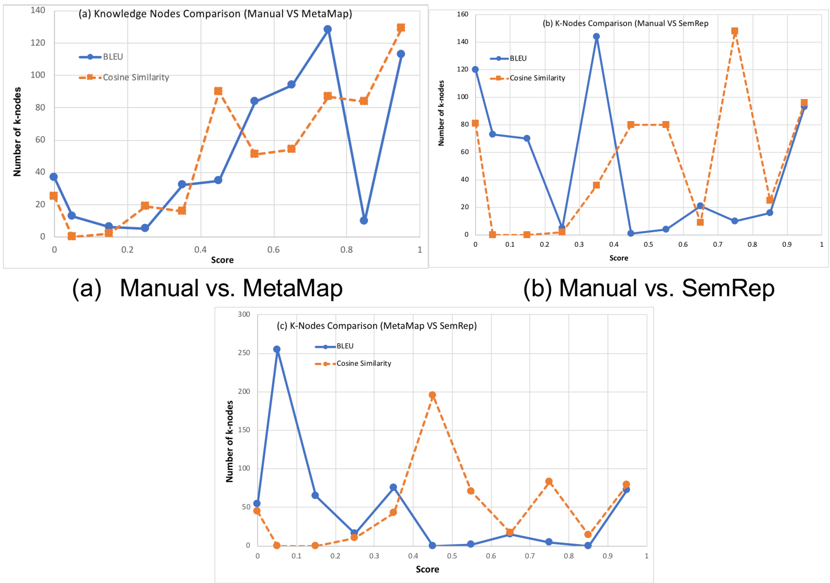
5 Evaluation

To compare the manual annotation results to those generated by automatic tools, we used MetaMap and SemRep to extract k-nodes and relations and applied the BLEU and cosine similarity algorithms to calculate the similarity scores. The average similarity scores in Table 5 show that MetaMap and manual methods in k-node recognition has a higher degree of similarity compared to other two pairs. Figure 1 visualizes the comparisons between each pair of three methods. This is also illustrated by Figure 1 (a) where more k-nodes had higher scores for both measures than the other two pairs, as Figure 1 (b) and (c) show that the scores fluctuated widely and the scores from BLEU and cosine similarity methods were not as consistent as that in Figure 1 (a).

Table 5. Similarity scores by using BLEU and cosine similarity measures for manual, MetaMap, and SemRep methods

	Manual vs. MetaMap	Manual vs. SemRep	MetaMap vs. SemRep
Average BLEU score	0.633	0.349	0.237
Average score for cosine similarity	0.685	0.585	0.556
Total number of k-nodes	557	557	557

We also evaluated the similarity between the k-nodes generated by three methods and matching UMLS concepts in each result set. Table 6 shows the total numbers of UMLS terms for each comparison as well as the average scores. Overall, UMLS matching had much lower averages across three comparing pairs than the scores in Table 5.



(a) Manual vs. MetaMap

(b) Manual vs. SemRep

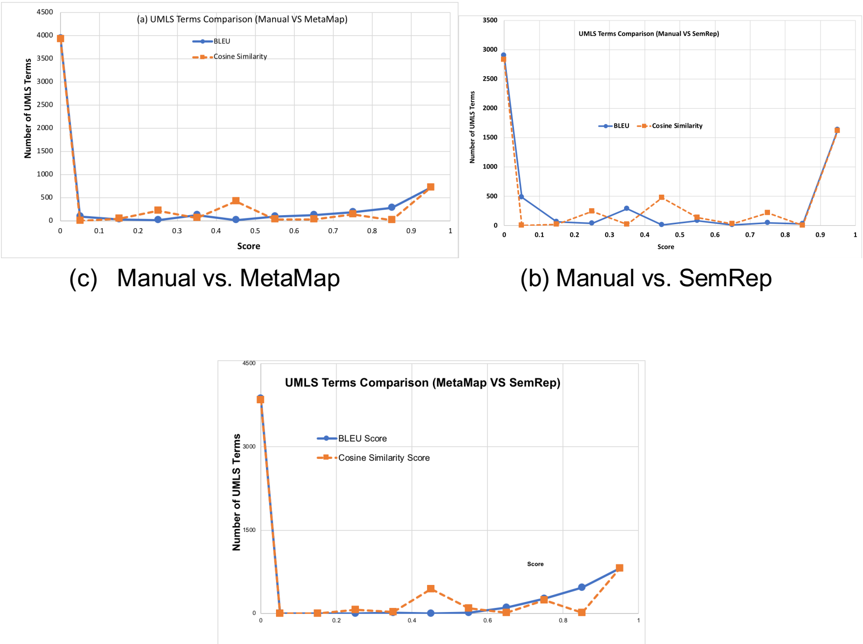
(b) MetaMap vs. SemRep

Fig. 1. Similarity evaluation scores for three k-node detection methods

It is worth pointing out that both BLEU and cosine similarity followed consistent patterns, that is, there was a large number of 0's and 1's in the scores. This suggests that a high number of k-nodes detected in the text did not have their counterparts in UMLS. This is consistent throughout the mapping between UMLS and each of the three k-node detection methods.

Table 6. Similarity scores for comparing UMLS concepts generated by manual, MetaMap, and SemRep methods

	Manual vs. MetaMap	Manual vs. SemRep	MetaMap vs. SemRep
Average BLEU score	0.227	0.337	0.269
Average score for cosine similarity	0.204	0.384	0.233
Total number of UMLS terms	5589	5589	5581
No matching for UMLS terms	168	299	438



(c) Manual vs. MetaMap (b) Manual vs. SemRep

(c) MetaMap vs. SemRep

Fig. 2. Similarity evaluation scores from both BLEU and Cosine similarity measures for three k-node detection methods

Discussion and Conclusion

knowledge node and relation recognition from full-text documents is highly challenging, yet critically important in the big data era. This paper is an attempt to examine different approaches in k-node and relation detection and hope to be able to offer some insights into the strengths and limitations of manual and automatic methods. Due to the small size of the data, the findings from this study are far from conclusive in addressing the three questions raised at the end of methods section. Nevertheless, we can learn a few things from the findings.

First, each of the manual and automatic k-node and relation detection methods has different areas of strengths and limitations. Manual method can offer fine granularity for k-nodes and relations, but it is slow and heavily relies on human coder's knowledge and analytical capability. MetaMap does a decent job in capturing k-nodes but lacks the capability for relation detection. SemRep can derive concepts and relations to put them in a three-part proposition format, but the k-nodes extracted by SemRep are essentially single words and the relations are strictly limited to what is available in UMLS Semantic Network. K-node and relation detection has much to be desired in the AI era.

Second, there is clearly a gap between effective knowledge representation from full-text documents and the tools. Here the tools include not only software but also codified rules and knowledge assertions. Although knowledge representation and knowledge base research have achieved a great success (e.g., the Cyc knowledge base and DBpedia), representing knowledge from full-text and codifying it for applications remains a challenging research field. Having upper-level KOS can be useful for knowledge representation from full-text, but they can be useful only at a coarse level; deeper and more refined knowledge representation for full-text requires more research on natural language processing and computational knowledge organization systems.

Finally, automatic relation detection is perhaps the crown in knowledge representation because one must first identify the k-nodes before relations among the nodes can be determined. Knowledge assertions in three-part proposition form (or simply, triples) has its root in AI. This three-part proposition is the foundation of modern KR that has been actualized in semantic web technologies – RDF, Linked Data, and Web Ontology Language (OWL). KR combining with semantic technologies have promises to revolutionize information retrieval, presentation, and use, but such promises have to meet the prerequisite – knowledge codified and structured as the content infrastructure upon which innovations in information retrieval, presentation, and use can be built. Our work on k-node and relation detection is a step toward this vision.

Representing knowledge in full-text documents is not a new research field, but with technology advances (semantic web, graph database, deep learning, etc.), this "old" research field is being injected fresh possibilities. Our future study will explore other tools such as NOBLE Coder on larger text collections.

References

Ahlers C. B., Fiszman, M., Demner-Fushman, D., Lang, F. M., Rindflesch, T. C. (2007) Extracting semantic predications from Medline citations for pharmacogenomics. *Pac Symp Biocomput. 2007; ():209-20.*

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46 (3): 175–185. doi:10.1080/00031305.1992.10475879.

Aronson, Alan R., and François-Michel Lang. "An Overview of MetaMap: Historical Perspective and Recent Advances." Journal of the American Medical Informatics Association 17, no. 3 (May 1, 2010): 229–36. doi:10.1136/jamia.2009.002733.

Brownlee, J. (2017). A gentle introduction to calculating the BLEU score for text in Python. https://machinelearningmastery.com/calculate-bleu-score-for-text-python/

Bushman, B., Anderson, D., & Fu, G. (2015). Transforming the Medical Subject Headings into linked data: Creating the authorized version of MeSH in RDF. *Journal of Library Metadata*, 15(3-4): 157-176.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is knowledge representation? *AI Magazine*, 14(1): 17-33.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science,* 41(6): 391–407. doi:10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9.

Divita G, Tse T, Roth L. Failure analysis of MetaMap transfer (MMTx). Medinfo. 2004;11(Pt 2):763–7.

Dobrokhotov, P. B., Goutte, C., Veuthey, A.-L., & Gaussier, E. (2003). Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. Bioinformatics, 19(Suppl. 1): i91-i94.

Hristovski D, Friedman C, Rindflesch TC, Peterlin B. (*2006).* Exploiting semantic relations for literature-based discovery. *AMIA Annu Symp Proc. 2006; ():349-53.*

Keikha, M., Khonsari, A., & Oroumchian, F. (2009). Rich document representation and classification: An analysis. *Knowledge-Based Systems*, 22(1): 67-71. doi:10.1016/j.knosys.2008.06.002

Kilicoglu, H., Rosemblat, G., Fiszman, M., & Rindflesch, T. C. (2011). Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics*, *12*, 486.

LC. (2018). LC Linked Data Service: Authorities and Vocabularies. http://id.loc.gov/

Liu, Y., Bill, R., Fiszman, M., Rindflesch, T., Pedersen, T., Melton, G. B., & Pakhomov, S. V. (2012). Using SemRep to label semantic relations extracted from clinical text. In *AMIA annual symposium proceedings* (Vol. 2012, p. 587). American Medical Informatics Association.

McCray, A. T. (2003). An upper-level ontology for the biomedical domain. *Comparative and Functional Genomics*, 4: 80-84. DOI: 10.1002/cfg.255

Mills, A. & Brickley, D. (2005). SKOS core guide: W3C working draft 2 November 2005. https://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/

NLM. (2018). Medical Subject Headings RDF. https://id.nlm.nih.gov/mesh/

NLM. (2016). Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/

Papineni, K., Koukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evation of machine translation. In: ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, July 07 - 12, 2002, pp. 311-318. Stroudsburg, PA: Association for Computational Linguistics.

Perl, Y., Geller, J., Halper, M., Ochs, C., Zheng, L., Kapusnik-Uner, J. (2016). Introducing the Big Knowledge to Use (BK2U) challenge. Annals of the New York Academy of Sciences, 1387(1): 12-24. https://doi.org/10.1111/nyas.13225

Qin, J., & Zou, N. (2017). Structures and relations of knowledge nodes: Exploring a knowledge network of disease from precision medicine research publications. In: *iConference 2017 Proceedings* (pp. 56–65). https://doi.org/10.9776/17009

Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of biomedical informatics*, *36*(6), 462-477.

Salton, G., Wong, A., & Yang, C. S. (1975), A vector space model for automatic indexing. *Communications of the ACM*, 18(11): 613–620.

Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P., & Musen, M. A. (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics,* 10(Suppl 9): S14. doi:10.1186/1471-2105-10-S9-S14.

Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., & Jacobson, R. S. (2016). NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics,* 17: 32. doi:10.1186/s12859-015-0871-y.

Van Harmelen, F., Lifschitz, V., & Porter, B. (2008). *Handbook of Knowledge Representation.* Oxford, UK: Elsevier.

Zhong, M., & Huang, X. (2006, August). Concept-based biomedical text retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 723-724). ACM.