Application of IBM Watson in the Medical Field

Utkarsh Kumar
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA 56267
kumar375@morris.umn.edu

ABSTRACT

There is a vast amount of data generated in medical research. They promise insights and breakthroughs for researches, but are a challenge to analyze and comprehend. IBM Watson is a cognitive computing tool designed to harness volumes of data, understand their various formats and make novel connections. This paper will look at a study using Watson to identify gene mutations in ALS patients and how it can further the pace of research.

Keywords

IBM Watson, Big Data, Natural Language Processing

1. INTRODUCTION

Innovation in medical science research is becoming costlier every year. A new drug is estimated to cost up to $2 billion and a decade of investment or more [3]. Out of potential drug candidates, around 80% fail to gain the approval of FDA with the most common reasons being lack of safe efficient results as well as poor dosage selection. Therefore, new drugs must be significantly more effective and safe. Furthermore researchers are often pressured to minimize time and expenses to meet deadlines such as in the case of COVID-19.

Fortunately, there are extensive sets of published research available to make informed decisions such as choosing the best drug candidates to continue research with. These informed decisions are based on complex human cognitive functions such as learning, reasoning and inference. However, human cognition is limited in *scalability*.

As of 2018, there are more than 28 million abstracts [2] in 5000+ journals in the MEDLINE corpus alone with more than 1.8 million published annually [3]. In contrast, the average researcher reads 250 to 300 articles in a given year [3]. This makes it impractical for a researcher to keep up with the latest developments and be fully informed about recent evidence that may be related to their study.

One tool that could help researchers is cognitive computing. Cognitive computing combines the capabilities of AI to read, reason and learn grouped with multiple technologies to provide a holistic solution to data challenges. IBM Watson is an application of cognitive computing that has been utilized in the medical field to assist researchers.

The rest of the paper will be as follows. Section 2 details the sources and various structures of data in the medical field as well as challenges they present for Watson. Section 3 breaks down and explain Watson's cognitive analytics in four subsequent procedures. Section 4 explains how Watson summarizes those analytics to output numerical values to researchers. Section 5 discusses Watson's performance as presented in a comprehensive study [1]. Section 6 discusses the challenge of bias followed by my conclusions.

2. UNDERSTANDING DATA

To further understand how Watson can help researchers, we need to understand data formats present in published medical research. Some provide challenges to Watson while others make Watson's computations easier.

One challenge of understanding text in medical research papers is the numerous representations or synonyms of a term in chemical nomenclature. For example, Valium is the brand name of the generic drug Diazepam. Apart from those two names, there are 148 other names Valium could be refereed by in a paper. Sections 3 and 5 include examples of chemical nomenclature and how Watson handles them in.

Another data format are Medical Subject Headings or *MeSH*. It is a series of vocabulary terms, manually curated by the Nation Library of Medicine [3], assigned to articles to help index, catalog and search health information across major medical database and catalogs. Examples of how Watson uses MeSH are given in Section 3.3.

3. HOW WATSON WORKS

The very first step in replicating human learning is the observation of data. Humans observe data by reading, listening, watching and other sensory inputs. They also use their pre-existing knowledge to understand the context for these observations. The new observations are then added to the set of pre-existing knowledge. Similarly, in order to make observations Watson must first have volumes of pre-existing or foundational knowledge from its prior learning.

3.1 Foundational Knowledge

The foundational knowledge used by Watson is aggregated by IBM from external, public, licensed and private sources of content. The data is then stored in a single repository called the Watson corpus. Just like humans heading out of college have collected knowledge on their specific area of study, a unique corpus is established for Watson depending on the domain it's applied to. The corpus for Watson application in the field of law or finance would have a uniquely different

corpus and foundational knowledge compared to Watson's corpus applied to medicine. To learn domain-specific knowledge, a corpus needs to contain dictionaries of names and synonyms of *entity types*.

For life science, the key concepts that Watson is trained on are genes, drugs, diseases, symptoms, and chemicals. These are referred to as entity types and any individual gene, drug, etc is referred to as a single *entity*. The foundational knowledge provided by the corpus might contain various information about each entity such as the list of proteins associated with each gene or approval status of various drugs. Furthermore, there are several dictionaries of entities and their synonyms allowing Watson to recognize multiple representation of entities.

Once Watson has the relevant foundational knowledge, it extracts the key concepts through a set of annotators. A chemical structure annotator is able to extract chemical names and convert them to unique chemical structures while a gene or protein annotator can extract gene and protein names and resolve to a unique gene identity. Annotators can also identify the relationships among genes, drugs and diseases [2]. The following sections will explain how these operations are performed using examples Watson would encounter in published medical research.

3.2 Named Entity Recognition

The step after observation is recognition. Humans also understand knowledge based on the context it's presented in. Most people don't know what '1,3,7-trimethylpurine-2,6-dione' or "CHEMBL113" is, but if they read "...the oral administration of CHEMBL113 was observed to..." in a paragraph which previously discussed the effects of caffeine, they could recognize CHEMBL113 is potentially a compound identifier for caffeine. Recognition of entities found in research papers is performed through *rule-based* approach.

The rule-based approach to extract compounds is based on using dictionaries of compound names and synonyms as well as context rules. Common names like 'caffeine' are limited in number, so they can be identified if they are provided in a dictionary as a part of foundational knowledge. Similarly, compound names likes 'CHEMBL113" can be looked for in the dictionary of synonyms mentioned in Section 3.2 or identified with regular expression.

Context rules prevent a compound string such as "nitric oxide" from being extracted if it occurs in the context of "nitric oxide synthase" that suggest the noun in question is a more specific noun denoted by a longer, overlapping phrase. Another example of context rules is the identification of contextual abbreviations and acronyms. Acronyms cannot be reliably identified based on dictionaries due to the vast number of words that have the same acronyms and the meaning of an acronym can change from one document to another. In the medical domain, alone 'DA' might refer to 'descending aorta', 'digital angiography', 'diabetic acidosis' or simply 'dopamine'. To resolve the meaning of acronyms, Watson identifies an acronym the first time it's defined in a paper and then persists that information across that document and only that document. This results in other uses of the same acronym being consistently interpreted across a single document, but being discarded when Watson analyzes the next document in the corpus.

Canonical gene names	DNM1L	DARPK2	DENR
PINK1	0.192771	0.036145	0
Parkin	0.37671	0	0
promote	0.000680	0	0.00068
Drp1-dependent	0.222222	0	0
mitochondrial	0.017527	0.008238	0.002665
fission	0.089744	0.028340	0.016869
Animals	0.000432	0.000235	0.000135
COS cells	0.002001	0.000858	0.000250
Dynamins/metabolism	0.106719	0.079051	0.015810
Protein Binding	0.000386	0.000303	0.000052
Signal Transduction	0.000127	0.000381	0.049282
...			
Total	0.730355	0.200110	0.049282

Table 1: Predictive context model scores between possible canonical forms of Drp1 and each context word, bolded, from the example sentence as well as MeSH terms, shown in italics, from example document. The gene with the highest score is most likely to be the canonical form of the ambiguous gene Drp1. [3]

3.3 Named Entity Resolution

The recognition step is followed by the interpretation step. After Watson has recognized the relevant strings, it performs named entity resolution which is the process of mapping the recognized strings to an entity. Each entity has a *canonical form* which is used to store and query all data associated with that entity. This allows any single entity and the various string referring to it across a set of documents to be grouped and identified in a search query where the user might use any these representations. Watson uses a hybrid approach to named entity resolution with different normalization techniques for different types of entities. As an example, I will cover the normalization technique used for two of these types; chemical compounds and genes.

3.3.1 Chemical Compound Normalization

In compound normalization, each compound's canonical form is a simplified molecular input line entry system or *SMILES* string. A compound string like "1,3,7-trimethylpurine-2,6-dione" is converted into a 2D chemical structure using *name-to-structure*, N2S, software which systematically decomposes the compound using chemical name formats and internal dictionary lookups [3]. Strings that fall outside the scope of N2S dictionary are queried for in Watson's foundation knowledge. There are also numerous string modifications such as removal of quotations and parentheses, as well as spelling correction performed throughout the process of dictionary lookups.

Part A of Figure 1 shows the 2D representation of chemical compounds generated by N2S. Part B breaks the cycles in the graph, so the resulting structure is a tree. The edges of each cycle are numbered, so the chemical representation can be reconstructed. In part C, each of the colored branches off the main green branch is traversed first in a depth-first traversal of the entire tree. Part D shows the generation of a SMILES string by printing the nodes in the order they are traversed. Carbon atoms are not labelled in chemical diagrams used here, but they are represented in the SMILES string. See [7] for additional details.

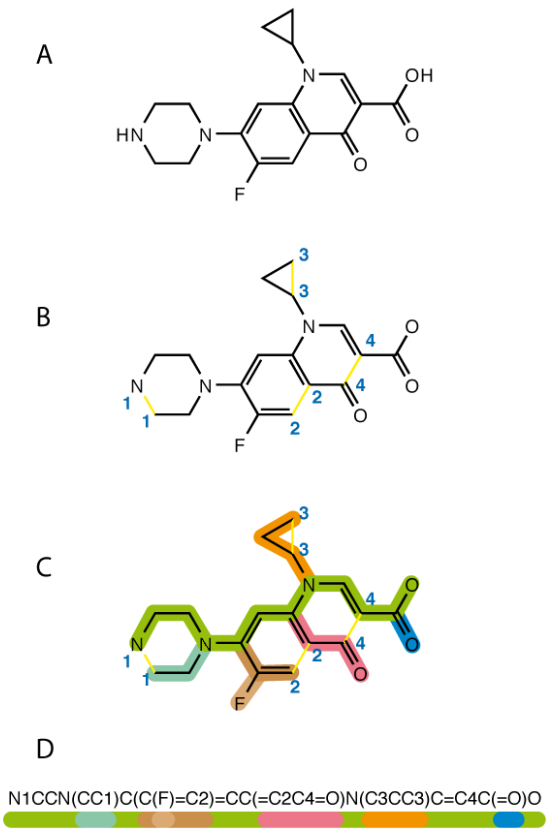

Figure 1: SMILES generation algorithm for Ciprofloxacin: break cycles, then write as branches off a main backbone [7].

3.3.2 *Gene Normalization*

The canonical form of a gene is typically stated in the foundational knowledge and often dictionary lookup can be used to map a synonym of a gene with that gene's canonical name. However, there are quite a few cases of ambiguous reference to a gene found in literature, therefore gene normalization process often involves building a predictive context model using other words surrounding the extracted string, referred to as context words, as well as the metadata of the document. An example of metadata would be the MeSH mentioned in Section 2.

While analyzing a research paper, Watson could comes across a sentence like: "We show that PINK1 and Parkin promote Drp1-dependent mitochondrial fission by mechanisms that are at least in part independent". Watson recognizes *Drp1* as an entity, but cannot map it a canonical form. This means Drp1 is an ambiguous gene name that could refer to various gene canonical names provided in Watson's foundational knowledge.

Table 1 shows the score describing how often each context word or MeSH term appears together with a gene synonym that is normalized to the corresponding canonical name. Higher scores indicate higher probability of that gene name being the canonical form of the ambiguous gene. In this case, Drp1 is normalized to the canonical name with the highest score: DNM1L.

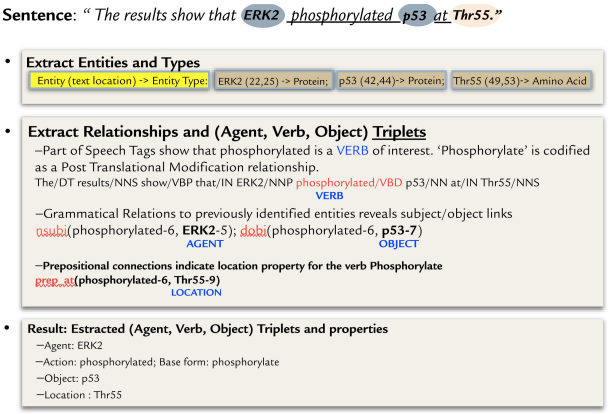

Figure 2: An example of Watson extracting relationships between terms from scientific literature [2].

3.4 Semantic Relationship Extraction

The last step is extraction of relationships. A relationship is generally defined as two distinct entities, an agent and a target/object, linked through a domain relevant verb, called a *trigger word* occurring in the same sentence. The list of trigger words are curated from multiple domain-specific databases as well as user feedback. In the sentence: "The results show that ERK2 phosporylated p53.", the trigger word would be 'phosporylated' with 'ERK2' being the agent and 'p53' being the target/object. Just like entities, the trigger words are also normalized so 'phosporylated' is normalized to 'phosporylate'. Trigger words that are more general such as 'bring' and 'overlap' are normalized to high-level relationships such as 'association'.

Some relationships also have a *residue*. If we expand the previous example sentence to "The results show that ERK2 phosporylated p53 at Thr55." an extra argument, 'Thr55' is identified as the residue, specifically a location, and normalized as threonine at position 55. The extraction of semantic relationship is shown in Figure 2.

Relationship consisting of non-entity nouns are captured as well. For a phrase such as "smoking increases the risk of lung cancer", Watson captures the relationship between smoking and lung cancer even though smoking is not recognized as an entity.

3.4.1 *Document Vectors*

At this point, Watson has the canonical form of all entities appearing in any of the documents it has analyzed. Watson then looks at how frequently these entities appear in a single document and creates a *document vector*. Entities that don't appear in a document have a frequency value of zero in its respective document vector. Once every document is represented as a document vector, those vectors can be averaged to create an *average document vector* which encodes the average frequency of each entity across all documents.

3.4.2 *Entity Vector*

Watson also represents each entity with an *entity vector*. An entity vector contains the average frequency of all entities across *only* the documents containing the entity being represented as an entity vector. Thus an entity vector represents that entity's *literature distance* to every other entity.

Algorithm 1: Create an n-ary similarity tree from a set of entities, based on [5].

Input: entities, n
Output: n-ary similarity tree

```
1  mostTypicalFV = average(entities)
2  root=closestTo FV(entities, mostTypical FV)
3  entities.remove(root)
4  candidates = root
5  while not entitites.isEmpty() do
6  |  (e, c) = closestPairs(entities, candidates)
7  |  c.addChild(e)
8  |  if c.numChildren() == n then
9  |  |  candidates.remove(c)
10 |  end
11 |  candidates.add(e)
12 |  entities.remove(e)
13 end
14 return root
```

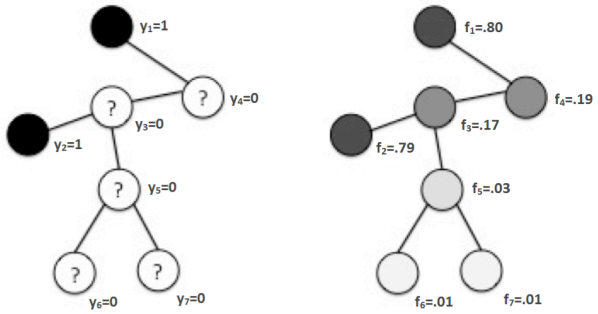

Figure 3: Example of graph diffusion on a similarity tree. (Left) Dark nodes, labeled 1, represent known entities. Light nodes, labeled 0, represent candidate entities. (Right) The darker nodes with GD values closer to 1 indicate more information in a node than the lighter nodes with values closer to 0. [5]

4. RANKING

To properly rank entities based on their similarity to another entity or a set of entities, Watson first creates a *similarity tree* and then applies a graph diffusion technique it. Each node in the similarity tree represent an entity and an edge represents the relationships between the two entities attached to that edge. There are two parts to creating the similarity tree: choosing an entity as the root node and generating the rest of the tree from that root node.

4.1 Generating the tree

Algorithm 1 shows the pseudocode for creating the similarity tree. The algorithm takes the set of all entities as well as an integer n as inputs. The first step is choosing an entity as the root node of the tree. To do this, Watson compares which entity vector is the closest to the average document vector. That entity is the most typical entity and the most reflective of the relationship between different entities and their frequencies for an average document [5]. Lines 3 and 4 show the removal of the root entity from the *entities set* and it's placement in the *candidates set*. Out of the entities set, the closest entity to the root node, based on literature distance, is added as its child node. This also results in the newly added node being removed from the entities set and added to the candidates set. New nodes are added to the tree by looking at which entity from the candidate set, already on the tree, is the closest to an entity from the entities set.

There is a check, shown on lines 8-10, to make sure no nodes have greater than n number of child nodes. For Watson, n=10 was chosen through cross-validation. This balances the tree between the extremes of having all other entities as the child of the root node or having a really tall tree with every node only having 2-4 children at most. The similarity tree is considered finished, shown on line 5, when there are no entities left in the entities set. [5]

4.2 Graph Diffusion

Once the similarity tree is built, we would like to know which entities, out of a candidate set, are most likely related to some known set of entities. To do this, Watson labels each node with 1 or 0. The value is 1 if that entity is in the known set and 0 if it's in the candidate set. Figure 3 shows an example of a graph diffusion algorithm applied to measure the flow of information between the nodes. The left graph in the figure visually represents nodes with value 1 as black and nodes with value 0 with a question mark.

The graph diffusion algorithm is heavily mathematical [5], but it can be thought of as observing heat transfer along metal pipes. After graph diffusion is applied to the similarity tree, the nodes closely surrounded by one or more known entities tend to have a greater amount of heat or information, which diffused from known entities, than nodes located at a farther distance from known entities.

The right graph in figure 3 shows information diffused from f_1, f_2 to f_3, f_4 and through them to the rest of the graph. The darker nodes, f_3 and f_4, have higher graph diffusion (GD) scores than lighter nodes $f_5 - f_7$, suggesting f_3 and f_4 are more similar to f_1 and f_2 than the other nodes.

The similarity tree in figure 3 is quite small. In practice these trees may have thousands of entities including dozens of known entities labeled as 1 in the tree.

5. IDENTIFYING ALS MUTATIONS

In a 2017 study, *Bakkar et al* [1] used Watson to identify potential candidates for RNA-binding proteins altered in ALS, described in Section 5.1. I will describe their usage of Watson and the results without going in depth into biological aspects of the study.

5.1 Background

ALS is a disease that affects nerve cells in the brain and spinal cord causing loss of muscle control. There are no effective treatments, however numerous RNA binding proteins, or *RBP*, have been shown to alter in ALS. There are at least 1,542 RBP-encoding genes in the human genome, 11 of which have shown to have a mutation causing a familiar, genetically inherited, form of ALS. There are also 6 other RBPs shown to be altered in ALS patients, but the gene producing them has not yet been linked to any known mutation that causes ALS. These still make less than 1% of RBPs studied and linked to ALS. It has been hypothesized that additional RBPs contribute to ALS and *Bakar et al* used Watson to predict potential candidates to study.

Protein	Rank
TARDBP	1
FUS	5
SETX	11
MATR3	12
TAF15	13
ATXN2	*21*
HRNPA2B1	*60*
ARHGEF28	*61*
HNRNPA1	106
GLE1	107
ANG	713

Table 2: Rank of known RBPs when removed from known set and placed into the candidate set. RBPs in bold ranked in the first 15 places out of 1,468 while another 3, italicized, ranked in the top 4.1% of candidates. Data from [1].

Since Watson uses text-based information from published research, the researchers could only use the 1,478 RBPs mentioned in at least one abstract published prior to 2016 in this study rather than 1,542 they knew existed.

5.2 Validating Watson

The researchers performed a leave-one-out cross validation (LOOCV) where an algorithm is applied multiple times with a different item being moved from the training set into the testing set to test the accuracy of the model. Watson applied the graph diffusion algorithm 11 times with a different RBP with known gene mutation placed into the candidate set alongside the other 1,478 RBPs each time. If the model is accurate, then the RBP placed into the candidate set should rank high based on the model built from the other 10 known RBPs. Indeed, the results of the LOOCV showed 5 of the 11 RBPs, bolded in table 2, ranking in top 15 out of 1,478 RBPs with three more ranking in the top 4.%1 italicized in Table 2.

The LOOCV also provided a point of reference for where to expect relevant RBPs. The results, provided in Table 2, can be extrapolated to determine that since 10 out of the 11 known RBPs ranked within top 8% of all RBPs, approximately 90% of RBPs possibly altered in ALS will appear in the top 8% of the rankings in subsequent analysis. [1]

5.3 Retrospective Analysis

The researchers then performed a retrospective analysis where the corpus of data analyzed by Watson was restricted to literature published up to the end of 2012. Only the 8 known RBPs with mutations linked to ALS in 2012, rather than the 11 known in 2017, were provided in the positive known set. Out of the currently known 1,478 RBPs being used in the candidate set, only 1,439 were mentioned at least once in the MEDLINE corpus of abstracts up to the end of 2012 thus those 1,439 RBPs were chosen as the candidate set for the retrospective analysis.

Table 3 shows Watson ranking of RBPs linked to ALS. MATR3 was identified as the top candidate while ARHGEF28 and GLE1 are ranked 89 and 165, respectively. Two of the six other RBPs, RBM45 and hnRNPA3, shown to have alteration but not linked to a mutation were also ranked highly with rank 8 and rank 45 respectively.

Candidate gene set	Score (GD)	Rank
MATR3	0.00204078	1
NUPL2	0.00181635	2
SRSF2	0.0017781	3
...		
hnRNPA3	0.00154361	8
RBM45	7.79E-04	43
ARHGEF28	3.95E-04	89
GLE1	3.85E-04	165

Table 3: Result of retrospective study. Watson ranked each gene based on semantic similarity of the candidate to the 8 known gene. The bold genes have been linked to ALS. Data from [1].

This retrospective analysis demonstrates that Watson, when given only literature published up to the end of 2012, could identify in top 11% of potential candidates every RBP linked to a mutation in ALS that would be found in the next 4 years between 2013 and 2017. If this technology was available and used in 2012, the researchers would have been able to identify those RBPs even sooner.

5.4 Prospective Analysis

Once the retrospective analysis established the performance capability of the model, the researchers performed a prospective analysis. 1,478 RBPs, 39 more than the retrospective analysis, were mentioned at least once in MEDLINE abstracts prior to 2016 and used as the candidate set for the prospective analysis. The known set includes all 11 known RBPs, identified prior to 2016, shown to mutate in ALS.

Only two proteins, underlined in Table 4, of the top ten candidates genes ranked by Watson had previously shown to be altered in ALS patients. Validation studies were performed with the other eight as positive control to see if they are altered in ALS. For negative control, three RBPs from the bottom of the the rankings were chosen with expectation of not seeing any alteration in those proteins.

Four different biological tests were carried out on these RBPs to check for alternation. These tests are purely biological so I won't cover them. It is significant that an RBP had to show statically significant difference between ALS and controls groups in at least two of these biological tests to be considered valid.

The tests' results showed that five out of the eight RBPs previously unlinked to ALS showed significant alterations. No alternation was found in the three RBPs from the bottom of Watson rankings indicating they are, as expected, not linked to ALS. Watson guided researchers to further examine 8 candidates out of 1,478, and five RBPs never linked to ALS before were discovered as a result.

6. THE CHALLENGE OF BIAS

As mentioned in section 5.1, the study could only use 1,478 RBPs mentioned in online published abstracts. This indicates that remaining 64 RBPs hadn't had the same level of research or discussion around them. Further research using Watson that builds only upon the results of *Bakkar et al.* [1] will similarly lack online published data on those 64 RBPs. If the usage of Watson becomes widespread, there is a danger of ignoring data represented in formats incompatible with Watson.

Candidate gene set	Score (GD)	Rank
hnRNPU	0.002914	1
SYNCRIP	0.002747	2
RBM45	0.002680	3
RBMS3	0.002494	4
SRSF2	0.002459	5
hnRNPH2	0.002255	6
NUPL2	0.002152	7
CAPRIN1	0.002109	8
RBM6	0.001915	9
MTHFSD	0.001910	10

Table 4: Top 10 candidates after the prospective study. Only the two underlined proteins, RBM45 and MTHFSD, have shown prior links to ALS. Data from [1].

Additionally, bias inherently present in published research can be further propagated by Watson. Especially in the medical field, the patient's race and ethnicity can be important predictors of trial outcomes, but only a small fraction of studies include the race or ethnicity of patients. Among trials that report demographic race or ethnicity data, the inclusion of minority patients is substantially lower expected based on the census demographics leading to decreased generalizability of trial conclusions across clinical populations. [4, 6].

The challenge doesn't necessarily needs to be addressed by Watson, but at least acknowledged that Watson's has potential to increase propagation of bias in the field.

7. CONCLUSIONS

Despite the challenges mentioned in the previous section, Watson is a powerful cognitive computational tool for analyzing published literature. As shown in Bakkar et al. [1], its application can lead to better selection of candidates for further examination significantly accelerating the pace and accuracy of research in the medical field. With further improvements in the field of cognitive computation as well as the exponential increase of data, we can expect to see widespread adoption of Watson and other similar cognitive computing tools by researchers looking to overcome the scalability limitations of human cognition.

Acknowledgments

I would like to thank Nic McPhee, Elena Machkasova, Ariel Cordes and the students of Senior Seminar Fall 2020 for their advice and feedback.

8. REFERENCES

[1] N. Bakkar, T. Kovalik, I. Lorenzini, S. Spangler, A. Lacoste, K. Sponaugle, P. Ferrante, E. Argentinis, R. Sattler, R. Bowser, and et al. Artificial intelligence in neurodegenerative disease research: Use of IBM Watson to identify additional RNA-binding proteins altered in Amyotrophic Lateral Sclerosis. *Acta Neuropathologica*, 135(2):227–247, 2017.

[2] Y. Chen, J. Elenee Argentinis, and G. Weber. IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research. *Clinical Therapeutics*, 38(4):688 – 701, 2016.

[3] R. L. Martin, D. Martinez Iraola, E. Louie, D. Pierce, B. A. Tagtow, J. J. Labrie, and P. G. Abrahamson. Hybrid natural language processing for high-performance patent and literature mining in IBM Watson for drug discovery. *IBM Journal of Research and Development*, 62(6):8:1–8:12, 2018.

[4] J. S. Somerson, M. Bhandari, C. T. Vaughan, C. S. Smith, and B. A. Zelle. Lack of diversity in orthopaedic trials conducted in the United States. *JBJS*, 96(7), 2014.

[5] S. Spangler, A. D. Wilkins, B. J. Bachman, M. Nagarajan, T. Dayaram, P. Haas, S. Regenbogen, C. R. Pickering, A. Comer, J. N. Myers, I. Stanoi, L. Kato, A. Lelescu, J. J. Labrie, N. Parikh, A. M. Lisewski, L. Donehower, Y. Chen, and O. Lichtarge. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1877–1886, New York, NY, USA, 2014. Association for Computing Machinery.

[6] A. Westervelt. The medical research gender gap: how excluding women from clinical trials is hurting our health. `https://www.theguardian.com/lifeandstyle/2015/apr/30/fda-clinical-trials-gender-gap-epa-nih-institute-of-medicine-cardiovascular-disease`, Apr 2015.

[7] Wikipedia. Simplified molecular-input line-entry system — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Simplified%20molecular-input%20line-entry%20system&oldid=980445922`, 2020.