Identification, characterization, and grounding of gradable terms in clinical text

Chaitanya Shivade[†], **Marie-Catherine de Marneffe**[§], **Eric Fosler-Lussier**[†], **Albert M. Lai**[*]
[†]Department of Computer Science and Engineering,
[§]Department of Linguistics,
[*]Department of Biomedical Informatics,
The Ohio State University, Columbus OH 43210, USA.
shivade@cse.ohio-state.edu, mcdm@ling.ohio-state.edu
fosler@cse.ohio-state.edu, albert.lai@osumc.edu

Abstract

Gradable adjectives are inherently vague and are used by clinicians to document medical interpretations (e.g., *severe* reaction, *mild* symptoms). We present a comprehensive study of gradable adjectives used in the clinical domain. We automatically identify gradable adjectives and demonstrate that they have a substantial presence in clinical text. Further, we show that there is a specific pattern associated with their usage, where certain medical concepts are more likely to be described using these adjectives than others. Interpretation of statements using such adjectives is a barrier in medical decision making. Therefore, we use a simple probabilistic model to ground their meaning based on their usage in context.

1 Introduction

Expressions used in a language are said to be vague if they do not convey a precise meaning. Sentences using vague expressions do not give rise to precise truth conditions (Kennedy, 2007). Consider the following sentence: "The patient was maintained on a *high* dose of insulin." Interpreting such statements is a problem since it is unclear what was the exact amount of insulin used. Gradability (Sapir, 1944; Lyons, 1977) is a semantic property that allows a word to describe the intensity of a measure in context, and thus enables comparative constructs. In the above example, the word *high* is said to be gradable since it conveys the meaning associated with the measure - amount.

Gradable adjectives inherently possess a degree of vagueness and are used in a language to express epistemic uncertainties (Kennedy, 2007; Frazier et al., 2008). While judgments are strong in extreme cases, there exist borderline cases, where it is difficult to ascribe an adjective. In the above example, some amounts of insulin would be considered as a high dose by all, other amounts would never be considered a high dose, but there is a middle range where it can be difficult for even experts to judge, if it is a high dose. This is because, different experts may have differing thresholds for what constitutes a high dose.

Broadly, gradable adjectives can be classified into two categories based on their interpretation as measure functions (Bartsch, 1975; Kennedy, 1999). Adjectives such as *tall, heavy, expensive* can be viewed as measurements that are clearly associated with a numerical quantity (height, weight, cost). In contrast, adjectives like *clever, beautiful, naive* are more complex and underspecified for the exact feature being measured. Gradable adjectives have been the focus of several recent studies (de Melo and Bansal, 2013; Ruppenhofer et al., 2014) in the NLP community. Gradablity is property not limited to adjectives and also extends to other parts of speech such as adverbs (Shivade et al., 2015; Ruppenhofer et al., 2015) (e.g., slightly, marginally), nouns (e.g., joy, euphoria), and also verbs (e.g., drizzling, pouring).

In this paper, we conduct a comprehensive study of gradable adjectives used in clinical text. Using a method proposed by Hatzivassiloglou and Wiebe (2000), we identify the gradable adjectives in our dataset of clinical notes. We found that these adjectives have a substantial presence (30%) in our data. Further, we show that there is a specific pattern in which gradable adjectives are used: some medical concepts are more likely to be modified by these adjectives than others. Finally, we focus on a specific subset of gradable adjectives associated with measurements of numerical quantities and demonstrate the use of a simple computational

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 17–26,
Berlin, Germany, August 12, 2016.

model to ground their meaning.

2 Dataset preparation

We used 58,880 clinical notes on Chronic Lymphocytic Leukemia (CLL), 2,652 notes on prostate cancer (PC) and 14,378 notes on Methicillin-resistant Staphylococcus Aureus (MRSA) representing three different cohorts from our institution as a corpus for our study. Thus we had a total of 75,910 notes with an average word count of 1,476 words per note. In addition, we also had access to 8,192 echocardiograms, which are cardiology reports mostly containing semi-structured data with few lines of free text (avg. word count = 64). All clinical notes were from adult patients collected for a period from 2005 to 2010 with necessary approval of the institutional review board at our institution.

These notes are written by healthcare professionals communicating different aspects of patient care and therefore correspond to different note types. For instance, "Progress Notes" are written by physicians documenting periodic developments in the condition of patients, their diagnosis, and treatment. "Operative Notes" are written by surgeons documenting the pre-operative diagnosis, description of the procedure, and the post-operative condition. Our corpus consists of notes belonging to 98 different note types. The name of each note type is mentioned in the first few lines of a templated document header and often has multiple lexical variations. For instance, a "Progress Note" can be an "Inpatient Progress Note" or an "Outpatient Progress Note." These names were manually normalized to 18 note types, and confirmed by a physician for correctness. Each note from our dataset was thus mapped to one of these normalized types.

Clinical notes have a typical structure: the content is often organized in sections (e.g., "History of Present Illness" followed by "Physical Examination" and ending with "Assessment and Plan"). The beginning of a section is formatted as distinct text with the section name in capital letters followed by a newline characted. We used a simple rule-based system to identify section headers and map the contents of a note to these sections. As with note types, section names also had multiple lexical variations (e.g., "Physical Examination" can be "Physical Exam" or "Physical Assessment" or simply "Exam"). Our corpus had 587 section names which were normalized to 17 note sections with a physician's approval.

3 Identification of gradable adjectives

First, we want to automatically identifiy gradable adjectives in our corpus. We reimplemented the method described in (Hatzivassiloglou and Wiebe, 2000), a log linear regression model that learns the weights associated with two features: 1) Number of times an adjective is used in comparative and superlative constructs, and 2) Number of times an adjective is modified by terms that intensify or diminish the semantic meaning of adjectives (mostly adverbs such as *very, little, somewhat,* etc. and a few nouns such as *bit,* etc.). Hatzivassiloglou and Wiebe (2000) manually created a list of 73 such terms. Their model was generated using the 1987 Wall Street Journal Corpus (Marcus et al., 1993) and tested on a hand curated gold standard dataset of 453 adjectives (235 gradable and 218 non-gradable) created using the Collins Birmingham University International Language Database dictionary, which is annotated for gradable and non-gradable adjectives.

We developed a logistic regression model with the two features described above. For the first feature, a morphology analysis component was developed to identify inflections of adjectives from their base form. This consisted of identifying adjectives in their comparative form using simple parts-of-speech tagging (Toutanova et al., 2003) and regular expression based rules. Although the test set used in (Hatzivassiloglou and Wiebe, 2000) is available, the list of 73 noun phrases and adverbial modifications is not. We therefore compiled this list using ten fold cross validation to capture the second feature. In each fold of training, we found all the adverbs and nouns modifying the gradable adjectives using the Stanford Dependency Parser (version 2.0.4) (de Marneffe et al., 2006). We determined the best subset by choosing an optimal threshold for the ($k = 81$) most frequent modifiers through cross validation. This gave us the second feature for gradability.

Although the method was developed on newswire text, we found that it worked surprisingly well for our clinical corpus. We trained the model on clinical notes and evaluated it on the test set published by Hatzivassiloglou and Wiebe (2000). Of the 453 adjectives in that gold standard test set, we found that 61 adjectives (e.g. *wealthy,*

Study	Corpus	Gradable	Non-gradable	Precision	Recall	F-Score
H & W(2000)	1987 WSJ	235	218	94.15	82.13	87.73
Our study	Clinical notes	217	175	99.51	84.32	91.34

Table 1: Performance of gradable adjective identification on the test set from Hatzivassiloglou and Wiebe (2000).

zesty) were not present in our corpus, resulting in a total of 392 adjectives (217 gradable and 175 non-gradable). Table 1 outlines (does not compare) the performance of classification in the two studies. Since the F-score of our model is reasonably high, we use it to identify the gradable adjectives in our corpus. In addition to the 392 adjectives present in the test set, the model identifies 1,709 gradable adjectives in our data. These were domain-specific words such as *therapeutic, retroperitoneal, edematous*, common adjectives such as *acute, febrile, gentle, pale*, and also some interesting compositions such as *well-nourished, low-normal*, and *near-complete*.

4 Usage characterization

Vagueness induced by gradable adjectives has been studied by researchers in the past. We want to investigate how frequently such language appears in clinical notes, and if there are certain situations where these terms are more likely to be used. In the following sections, we show that not only do gradable adjectives have a substantial presence in clinical text, but there is also a definite pattern in their usage.

4.1 Presence of gradable adjectives

Using the model described in the previous section, we found all gradable adjectives present in our corpus. The percentage of adjectives identified as gradable in the notes across the 18 normalized note types was calculated. This percentage is fairly consistent across different note types, $\mu = 30.85\%, \sigma = 4.9\%$.

In addition to examining the distribution of gradable adjectives across notes types, we performed a finer analysis by calculating their percentage across different sections in a note. The percentage of adjectives identified as gradable across the 17 normalized sections was calculated. Again, it if fairly consistent ($\mu = 31.45\%, \sigma = 6.2\%$) across different sections.

4.2 Usage pattern

In this section, we present statistics that characterize the usage of gradable adjectives in describing medical concepts of different semantic types in clinical notes. The Unified Medical Language System (UMLS) (Lindberg et al., 1993) is a repository of multiple biomedical vocabularies and standards, developed by the US National Library of Medicine. A major component of the UMLS is the Semantic Network which assigns a semantic type to every concept. A semantic type is a high-level category (e.g., "Sign or Symptom," "Pharmacological Substance," "Plant," "Enzyme") analogous to named-entity types and there are 133 such semantic types in the 2013AA version of the UMLS.

MetaMap (Aronson, 2001) is a program that can map words from free text documents to concepts from the UMLS. Using the Stanford Dependency Parser, we identified medical concepts that were modified by a gradable adjective in our corpus and looked up their semantic types. For example: in *extreme fatigue*, the gradable adjective *extreme* modifies the term *fatigue* which has the semantic type "Sign or Symptom," while in *severe stenosis*, the adjective *severe* modifies the term *stenosis* which has the semantic type "Disease or Syndrome."

We hypothesized that gradable adjectives modify certain nouns more often than others. In order to test this hypothesis, we calculated how often nouns of a particular semantic type are modified by gradable adjectives. These frequencies were calculated for the three sets of clinical notes corresponding to three different diagnoses (CLL, PC, and MRSA) in our corpus. Nouns from a certain semantic types were very frequently described using gradable adjectives (e.g., "Finding," "Therapeutic or Preventive Procedure," "Disease or Syndrome"), and hence had high frequency values in all three datasets. Similarly, nouns from a few semantic types were never described by gradable adjectives (e.g., "Reptiles," "Professional Society").

Dataset	CLL	PC	MRSA
CLL	1.00	0.93	0.90
PC	0.93	1.00	0.91
MRSA	0.90	0.91	1.00

Table 2: Spearman's Correlation between clinical notes for semantic type modification by gradable adjectives.

We confirmed this by sampling each dataset into five equal folds and repeating the frequency calculations. The observations for frequency variations were consistent for every fold across each dataset. We performed a simple add-one Laplace smoothing to account for low frequency semantic types across datasets. Since the size of the three datasets were significantly different, we normalized the frequencies by the sum of frequencies across all semantic types within each dataset. The normalized frequency values represent the probability of a semantic type being modified by gradable adjectives in a dataset. We computed the Spearman's correlation for these 133 probabilities across each pair of datasets and found that there was a high correlation between them (Table 2). This high correlation across all three diagnoses suggests a definite pattern for the usage of gradable adjectives in clinical text.

5 Probabilistic Modeling

Gradable adjectives are widely studied as implicit or explicit measurements of certain quantities (Bartsch, 1975; Kennedy, 1999). Moreover, they also participate in a scale. For example, the adjectives (*warm* $<$ *hot* $<$ *scorching*) represent a scalar relationship and implicitly measure temperature. While judgments to associate an adjective with extreme values are very strong, those for borderline cases are difficult. In the above example, certain values of temperature are definitely *warm* and others are definitely considered *hot* (and yet not scorching). But there is always a set of values in between which can be either *warm* or *hot*. In order to capture this intuition, we created a probabilistic model using Bayes rule:

$$P\left(grad|num\right) = \frac{P\left(num|grad\right) \cdot P(grad)}{P(num)} \quad (1)$$

where *grad* represents the gradable term and *num* the numerical value.

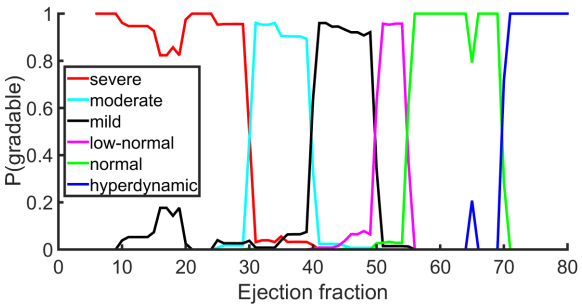

Figure 1: Probabilistic modeling of adjectives describing systolic function.

Clinicians frequently document their assessments for a patient along with evidence to support their claim, e.g., "`Mild anemia, Hgb 8.2`." This sentence has a medical concept "anemia" being described by a gradable adjective *mild* on the basis of the measurement of a numerical value - hemoglobin. For several medical concepts, we extracted using regular expressions, instances where an assessment for a medical concept was made using a gradable term, along with a numerical evidence to support the claim. Specifically, we indexed all sentences using Lucene and searched for ones containing the medical term (e.g. anemia) and the quantity of interest (e.g. hemoglobin). Finally, numerical values and adjectives were extracted using regular expressions. In the following subsections, we demonstrate that we can ground the meaning of gradable terms using the above model.

5.1 Systolic Function

Systolic function is a measure of how well the lower left pumping chamber of the heart sends blood to the rest of the body. It is measured using a numerical quantity called left ventricular ejection fraction (LVEF) which is documented in an echocardiogram. There is variation among physicians defining the precise threshold for a normal ejection fraction (Sanderson, 2007). While normal values range from 55 to 65, values less than 30 imply that the systolic function is severely compromised. We extracted LVEF values from the echocardiogram reports and their corresponding descriptions of systolic function. Posterior probabilities $P(gradable|LVEF)$ were calculated using equation (1) which resulted in a plot as shown in Figure 1.

From the 8,192 echocardiogram reports, we found six gradable adjectives in association with

LVEF values. While the adjectives *severe, mild* and *moderate* are associated with systolic dysfunction, the adjectives *low-normal, normal* and *hyperdynamic* are associated with systolic function. Although there is discussion in the clinical community regarding qualitative descriptions for ejection fraction (Radford, 2005), there is variation in these recommendations. Moreover, certain terms though used frequently (e.g. *low-normal*) are never a part of such guidelines.

An interesting observation can be made regarding Figure 1, drawing an analogy from the concept of WordNet *dumbbells* (Sheinman et al., 2012). A WordNet dumbbell is a representation involving an antonym pair (e.g. *small* and *large*) as two ends of a semantic scale with semantically similar adjectives arranged in a radial fashion around each adjective. The antonym acting as a centroid and its synonyms as members of a cluster represent words that most likely participate in the same scale. For example, the antonym pair (*small, large*) results in the dumbbell with clusters (*small, tiny, pocket-size, smallish*) and (*large, gigantic, monstrous, huge*) at the two ends. WordNet dumbbells have been used in the past (Sheinman et al., 2013; de Melo and Bansal, 2013) to group gradable adjectives belonging to the same scale. It can be seen that the analogous dumbbell consisting of (*severe, mild, moderate*) and (*low-normal, normal, hyperdynamic*) can be constructed using the modified terms systolic dysfunction and systolic function respectively.

The model captures essential aspects of gradability very well. The scalar relationships (*severe* $<$ *moderate* $<$ *mild*) and (*low-normal* $<$ *normal* $<$ *hyperdynamic*) can be inferred by imposing an order on the mean values for the posterior distributions of these adjectives. Strong judgments for extreme cases and uncertainty for borderline cases can be observed in the form of flat peaks for specific intervals and overlapping distributions for mid-range values.

5.2 Anemia

Hemoglobin is a protein in the red blood cells (RBCs) that contains iron and carries oxygen from the lungs to the rest of the body. Anemia is a blood disorder, operationally defined as a reduction in the hemoglobin content of blood caused by a decrease in the RBCs below a reference interval of healthy individuals. The range of normal

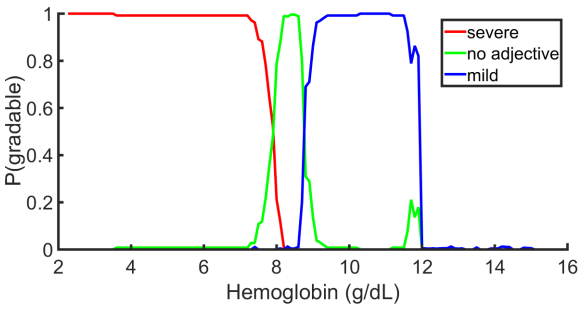

Figure 2: Probabilistic modeling of descriptions for anemia.

hemoglobin values for the laboratories at our institution is from 11.7 to 15.5. We found the two adjectives *severe* and *mild* to be most commonly used for describing anemia. A number of notes also mentioned anemia with no modifier at all. Figure 2 shows the posterior probabilities calculated for the three modifications of anemia: *mild, no adjective*, and *severe* using the model outlined in equation 1.

It is interesting to note that when physicians refer to anemia without an adjective, it is neither severe nor mild, and has a value in between. As with systolic function, we can infer the ordinal relationship (*severe anemia* $<$ *anemia* $<$ *mild anemia*), considering the mean values for the posterior distributions of these adjectives. Also, strong judgments for extreme values and uncertainty for borderline cases are evident through flat peaks and overlapping distributions respectively. We also found the adjective *moderate* being used in our data for describing anemia for hemoglobin values between *mild* and *severe*. However, it had few occurrences and hence we did not include *moderate* in our model. Other adjectives such as *significant*, *marked*, *slight* and *pernicious* were also found in the data but with low frequency counts.

5.3 Platelet count

Platelets (also known as thrombocytes) are colorless blood cells that help the process of blood clotting. There are about 150,000 to 450,000 platelet per microlilter of blood in the human body (Erkurt et al., 2012). While the condition resulting from a lower than normal platelet count is known as *thrombocytopenia*, the condition resulting from a higher than normal platelet count is referred to as *thrombocytosis*. Since the notion of *low* and *high* counts is gradable, we treat equivalent descriptions of thrombocytopenia and thrombocyto-

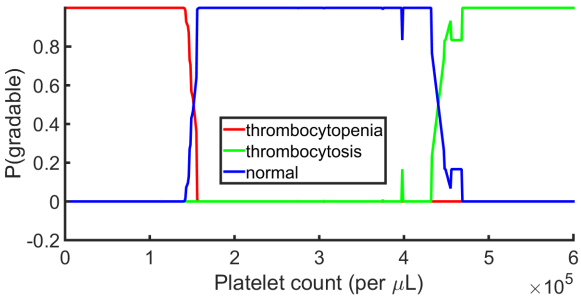

Figure 3: Probabilistic modeling of descriptions for variations in platelet count.

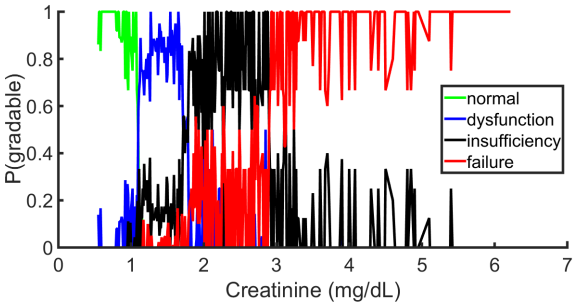

Figure 4: Probabilistic modeling of descriptions for variations in creatinine.

sis as gradable. In addition we also extracted instances of clinical notes where the platelet count was referred to as *normal*. Using these three descriptions, we applied the Bayes rule explained in Equation 1.

Figure 3 shows posterior probabilities calculated for these three descriptions of platelet count. As with previous examples, we can infer the ordinal relationship (*thrombocytopenia* $<$ *normal* $<$ *thrombocytosis*) by considering the mean values of their posterior distributions.

5.4 Renal Function

Creatinine is a chemical made by the body and is used to supply energy to the muscles. Creatinine is removed from the body by the kidneys and released through urine. If kidney function (or renal function) is not normal, creatinine level in the body increases (Israni and Kasiske, 2011). Abnormal renal function is referred to through different terminologies such as renal insufficiency, renal failure, and renal dysfunction. The vagueness introduced by the use of these gradable terms is also evident in clinical literature. Hsu and Chertow (2000) in their paper titled "Chronic renal confusion: insufficiency, failure, dysfunction, or disease" propose a set of laboratory values to classify patients as *mild*, *moderate* and *advanced* degrees of chronic renal insufficiency to "facilitate communication among nephrologists and other physicians and provide a framework for comparison of populations." It should be noted that linguistic ambiguity is not the only reason for this confusion and also has medical explanations which are beyond the scope of discussion of our work.

This problem was acknowledged by the medical community. More than 30 new definitions were proposed (Bellomo et al., 2004) and a new standard is now in place (Khwaja, 2012). However, our data is older (from 2005 to 2010) and has frequent occurrences of these terms. We extracted instances for the gradable terms "normal renal function," "renal failure," "renal insufficiency," "renal failure" and the corresponding creatinine values mentioned by physicians in the text. Further, we computed posterior probabilities for $P(gradable|creatinine)$ using our model (Figure 4). The range of normal creatinine values is between 0.60 to 1.10 for the laboratories at our institution. In comparison with other examples discussed so far, it can be seen from the plot that there is a greater confusion in the use of these terms. This is especially evident in the interval [2,3]. Again, this confirms with the property of uncertainty for borderline cases. However, an ordering (*normal* $<$ *dysnfunction* $<$ *insufficiency* $<$ *failure*) can still be inferred.

5.5 Evaluation

We evaluated the model to determine if it fits the data well. Using leave one out cross validation, we tested if the model was able to predict the adjective for a given numerical value. The gradable mentioned in each text extract was regarded as the gold standard prediction label. While creating a model, we ensured that there were at least three data points for each measurement value of the numerical quantity present in the data. This allowed us to compute priors for all values in the data. In practice, one would either need large amounts of data or employ smoothing (Kneser and Ney, 1995) to ensure prior calculations for all numerical values are possible. Accuracy is calculated across all gradable terms for each medical concept as described in previous sections (Table 3). The models achieve fairly high accuracies which demonstrates that our model fits the data well.

Medical concept	Number of data points	Accuracy (%)
Systolic function	10,201	90.4
Anemia	12,711	88.3
Platelet count	14,234	94.6
Renal function	16,309	74.8

Table 3: Evaluation of probabilistic models to predict gradable terms for numerical values in the data.

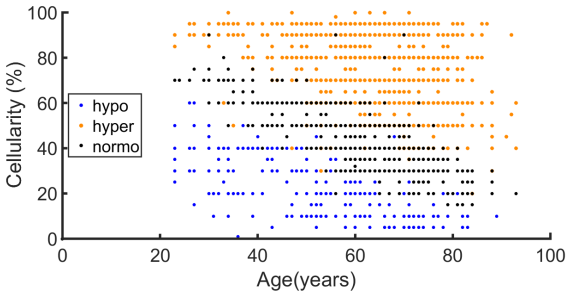

Figure 5: Dependency of gradable terms for BMC on age.

6 Limitations and future work

We illustrated through examples that gradable terms in clinical text can be effectively analyzed through data using a simple probabilistic model. The model is developed for cases where the use of gradable terms is dependent on a single numerical quantity. We included analysis of descriptions for heart function and kidney function. Similar analysis can be conducted for liver function which measures the amount of bilirubin in the body. Common tests such as body mass index, blood pressure and heart rate can also be analyzed in this way. Such a data-driven approach can help in creation of a standard terminology and avoid confusions (Hsu and Chertow, 2000).

However, context sensitivity is an important characteristic of gradable adjectives (Kennedy, 2007). Thus, "John is a *tall* boy" and "John is a *tall* basketball player" convey different meanings despite using the same gradable adjective for the same person (van Rooij, 2011). Similarly, the gradable description of a medical concept may not always be dependent on a single numerical quantity. For example, there is a slight variation in the upper limit of normal (ULN) values for creatinine with gender. The ULN for males is 1.3 while that for females is 1.2 at our institution. Similarly, the lower limit of normal for hemoglobin in males is 11.7 while that for females is 13.2. These variations are small in magnitude. However, this is a problem in cases where the dependency on other variables is much more pronounced. We illustrate this through an example.

Bone Marrow Cellularity (BMC) is the volume ratio of hematopoietic cells (blood cells that give rise to other blood cells) and fat. Pathologists perform a bone marrow analysis and use the three adjectives *hypocellular, normocellular*, and *hypercellular* to describe the sample. However, BMC is largely dependent on age of the patient. It is 100% for newborn infants and reduces with age in adults (Muschler et al., 2001). Therefore, the notion of hypocellular, normocellular, and hypercellular also varies with age. We extracted BMC values and associated adjectives from our data. Figure 5 shows the likelihood plot of BMC values against associated age of patients with three different colors for the adjectives *hypocellular, hypercellular*, and *normocellular*. Although the three gradable descriptions are linearly separable, $P(gradable|BMC)$ cannot be modeled using Equation 1, which ignores the age of the patient.

Time is a very common variable that often plays an important role in clinical assessments. This is most evident in blood sugar values for diabetic patients that vary with every hour depending on times of food consumption. Temporal adjectives are frequently found as descriptions of medical concepts. Some of the commonly found temporal adjectives in our data include *acute, chronic, recent, progressive, worsening, stable, persistent*, and *continued*.

Clinical notes are created and read by different individuals associated with the hospital. Vital decisions such as clinical trial recruitment, adherence to treatment guidelines, etc. are made by healthcare professionals based on their interpretation of these clinical narratives. Introducing automation in these processes is an active area of NLP research (Demner-Fushman et al., 2009). This decision making becomes challenging if language used in the clinical notes is vague and does not deliver a precise meaning. Our work is a small step to illustrate that gradability and its associated vagueness is an important aspect of clinical text which can be modeled through data. Creating a single model that can flexibly incorporate multiple variables and yet capture the properties of grad-

able adjectives can be an interesting line of research for the future.

7 Related Work

The phenomenon of adjectival modification in biomedical discourse has also been a subject of interest. Through empirical observations, Chute and Elkin (1997) classified frequent modifiers for medical concepts into two types: clinical modifiers (*e.g., chronic, severe, acute*) and administrative qualifiers (*e.g., history of, no evidence of, status post*). Bodenreider and Pakhomov (2003) extended this idea and compared adjectival modifications in biomedical literature and patient records. They found that while patient records contain markers for uncertainty (e.g., *possible, probable*) and non-specific symptoms (e.g., *low back pain, discomfort*), scientific articles are precise about attributes of organisms or age-groups (e.g., human, canine, neonatal).

Adjectives have been studied extensively in computational linguistics. WordNet (Fellbaum, 1998) classifies adjectives into two broad categories: descriptive and relational. Descriptive adjectives (e.g., *big* house, *heavy* bag) ascribe the value of an attribute to a noun, while relational adjectives (e.g., *atomic* bomb, *dental* hygiene) do not. Among the various distinctions between descriptive and relational adjectives, relational adjectives are typically not gradable (Fellbaum, 1998).

Although association between adjectives and numerical quantities has been a topic of research in some studies (Aramaki et al., 2007; Davidov and Rappoport, 2010; Iftene and Moruz, 2010), very few studies have investigated grounding the meaning of adjectives to numerical quantities. de Marneffe et al. (2010) investigated the problem of interpreting implied answers to yes/no questions when the response is not explicit. Specifically, they investigated question-answer pairs in which the question contains an adjective and the answer contains a numerical measure. For example, predicting the correct yes/no answer in (1) involves interpreting a numerical quantity (age) with respect to the gradable adjective *little*.

1. Q. Are your kids little?
   A. I have a 7 year-old and a 10 year-old.

The authors created logistic regression models for each adjective by querying the web with appropriate keywords ("little kids") and its antonyms ("not little kids"), so that both positive and negative instances can be learned.

Narisawa et al. (2013) explore a closely related problem of learning *numerical common sense* for the task of RTE in Japanese text. They study a broad set of cases that require semantic inference over numerical expressions. They query the web to gather instances of pairs of numerical quantities and corresponding contexts and propose two approaches. The distribution based approach concludes the numerical quantity to be *large* or *small* if it appears in the top or bottom five percent of the distribution generated for the numerical quantity and *normal* if it is in between. The cue-based approach relies on explicit textual cues (e.g., *as large as*, *only*) for associating a judgment about a numerical expression.

8 Conclusion

We empirically evaluated use of gradable adjectives in clinical documents. We reimplemented a previously published model for identifying gradable adjectives in newswire text and found that it performs surprisingly well with our clinical data. These adjectives have a substantial presence in clinical notes across multiple types of documents, written by different healthcare professionals. Analysis of the frequencies of these adjectives and their association with clinical concepts from UMLS revealed that there is a specific pattern for their usage. Finally, we showed that a simple Bayesian model can be used effectively to ground the meaning of gradable terms when they are used to describe medical concepts involving measurement of numerical quantities. Our data-driven approach can help in development of clinical standards in situations where there is a need to establish a precise relationship between adjectives and measurements.

Acknowledgements

We would like to thank Courtney Hebert and Kelly Regan for their help in this work. Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011116. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. 2007. UTH: SVM-based Semantic Relation Classification using Physical Sizes. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 464–467.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the Annual AMIA Symposium*, pages 17–21.

Renate Bartsch. 1975. The grammar of relative adjectives and comparison. In *Formal Aspects of Cognitive Processes*, pages 168–185. Springer.

Rinaldo Bellomo, Claudio Ronco, John A Kellum, Ravindra L Mehta, Paul Palevsky, and ADQI workgroup. 2004. Acute renal failure–definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Critical care*, 8(4):R204–R212.

Olivier Bodenreider and Serguei V. Pakhomov. 2003. Exploring adjectival modification in biomedical discourse across two genres. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*.

Christopher G. Chute and Peter L. Elkin. 1997. A clinically derived terminology: qualification to reduction. In *Proceedings of the AMIA Annual Fall Symposium*.

Dmitry Davidov and Ari Rappoport. 2010. Extraction and approximation of numerical attributes from the web. In *Proceedings of ACL*, pages 1308–1317.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC*, pages 449–454.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "Was it good? It was provocative". Learning the meaning of scalar adjectives. In *Proceedings of ACL*, pages 167–176.

Gerard de Melo and Mohit Bansal. 2013. Good, Great, Excellent: Global Inferences of Semantic Intensities. *Transactions of the Association of Computational Linguistics*, 1(July):279–290.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–72, Oct.

Mehmet Ali Erkurt, Emin Kaya, Ilhami Berber, Mustafa Koroglu, and Irfan Kuku. 2012. Thrombocytopenia in adults: review article. *Journal of Hematology*, 1(2-3):44–53.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Lyn Frazier, Charles Clifton, and Britta Stolterfoht. 2008. Scale structure: Processing minimum standard and maximum standard scalar adjectives. *Cognition*, 106(1):299–324.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 299–305.

Chi-yuan Hsu and Glenn M. Chertow. 2000. Chronic renal confusion: insufficiency, failure, dysfunction, or disease. *American Journal of Kidney Diseases*, 36(2):415–418.

Adrian Iftene and Mihai-Alex Moruz. 2010. UAIC participation at RTE-6. In *Proceedings of the Text Analysis Conference (TAC 10)*.

Ajay K. Israni and Bertram L. Kasiske. 2011. Laboratory assessment of kidney disease: glomerular filtration rate, urinalysis, and proteinuria. In *Brenner and Rector's The Kidney*, volume 9, pages 1585–619. Elsevier.

Christopher Kennedy. 1999. *Projecting the adjective: the syntax and semantics of gradability and comparison*. Routledge.

Christopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45, March.

Arif Khwaja. 2012. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4):c179–c184.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of ICASSP*, pages 181–184.

Donald A Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291.

John Lyons. 1977. Semantics (Volumes I & II). *Cambridge CUP*.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

George F. Muschler, Hironori Nitto, Cynthia A. Boehm, and Kirk A. Easley. 2001. Age-and gender-related changes in the cellularity of human bone marrow and the prevalence of osteoblastic progenitors. *Journal of Orthopaedic Research*, 19(1):117–125.

Katsuma Narisawa, Yotaro Watanabe, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. 2013. Is a 204 cm Man Tall or Small ? Acquisition of Numerical Common Sense from the Web. In *Proceedings of ACL*, pages 382–391.

Martha J. Radford. 2005. ACC/AHA Key Data Elements and Definitions for Measuring the Clinical Management and Outcomes of Patients With Chronic Heart Failure. *Journal of the American College of Cardiology*, 46(6):1179–1207, sep.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In *14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 117–122.

Josef Ruppenhofer, Jasper Brandes, Petra Steiner, and Michael Wiegand. 2015. Ordering adverbs by their scaling effect on adjective intensity. In *Proceedings of RANLP*, pages 545–554.

John E Sanderson. 2007. Heart failure with a normal ejection fraction. *Heart*, 93(2):155–158.

Edward Sapir. 1944. Grading, A Study in Semantics. *Philosophy of Science*, 11(2):93–116.

Vera Sheinman, Takenobu Tokunaga, Isaac Julien, Peter Schulam, and Christiane Fellbaum. 2012. Refining WordNet adjective dumbbells using intensity relations. In *Sixth International Global Wordnet Conference*, pages 330–337.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam, and Takenobu Tokunaga. 2013. Large, huge or gigantic? Identifying and encoding intensity relations among adjectives in WordNet. *Language Resources and Evaluation*, 47(3):797–816, January.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. 2015. Corpus-based discovery of semantic intensity scales. In *Proceedings of NAACL-HLT*, pages 483–493.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL-HLT*, pages 173–180. Association for Computational Linguistics.

Robert van Rooij. 2011. Vagueness and Linguistics. In Giuseppina Ronzitti, editor, *Vagueness: A Guide*, chapter Vagueness, pages 123–170. Springer Netherlands.