Targeting Explanations by Measuring Conceptual Complexity

Thomas Macaulay Ferguson[1,2][0000−0002−6494−1833]

[1] ILLC, University of Amsterdam, Amsterdam, The Netherlands
[2] Arché Research Centre, University of St. Andrews, St. Andrews, Scotland
tferguson@gradcenter.cuny.edu

Abstract. Explanations, *e.g.*, justifications or refutations of some truth or position, are not deployed in a vacuum. Rather, an explanation is generated to be consumed by *users*. Pragmatic considerations ensure that an explanation that a user cannot *understand* is as good as no explanation at all. Thus, an important consideration for the *fittingness* of an explanation for a particular user is whether she is likely to have sufficient conceptual maturity to grasp the concepts in play. In this discussion, we suggest that measuring conceptual complexity is critical to ensure that appropriate explanations are provided to given audiences. We also describe a general technique leveraged in assessing appropriateness of justifications in the Cyc knowledge base by appealing to *readability tests* applied to large natural language corpora.

Keywords: explainable AI · conceptual complexity · readability.

1 Introduction

Consider a knowledge system that is capable of generating justifications for its conclusions. These justifications are generated for a *reason*, namely, to supply a user with a deeper understanding of—and thereby, a deeper trust in—a conclusion. Clearly, there are many pragmatic constraints that the choice of user places on an explanation's adequacy. Natural language presentations of justifications are to be preferred over pseudocode-like presentations; in turn, natural language justifications should be provided in a language with which the user is familiar.

This discussion is oriented towards a different dimension: The constraint that a user should be able to *understand the concepts* or *subject-matter* to which a justification makes appeals. Now, a knowledge system employed by *e.g.* a conversational artificial intelligence may be responsible for interactions a wide range of interlocutors with a similarly wide range of intellectual development and specialization. A digital assistant in a household, therefore, may interact with a range of users including small children and adults with a wide range of education levels. It is obvious that a justification consumable by a PhD may not be consumable by every other member of the household.

To illustrate, suppose that a household digital assistant is given the (not uncommon) task of explaining simple facts about the world and is charged with providing answers to questions like:

Q: `Why does an arctic fox have a white coat in the winter?`

One might envision several explanations for this fact. One might encounter explanations ranging from the extremely fine-grained (paraphrased from [7]):

Exp1: `The arctic fox has a white coat in the winter because:`

- `During winter months in the arctic, there is less UV radiation than in summer months.`
- `Less UV radiation in an environment leads to the production of less` α`-Melanocyte-stimulating hormone (`α`-MSH).`
- `A high ratio of agouti protein to` α`-MSH leads to diminished melanocyte activity.`
- `In conjunction with the additional mass of the winter coat, this dilutes available pheomelanin, leading to a white coat.`

To the relatively simple, such as:

Exp2: `The arctic fox has a white coat in the winter because:`

- `White fur allows the arctic fox to blend in with its surroundings during the winter.`

It seems that **Exp1** and **Exp2** are equally *correct* responses to the query. Nevertheless, the two are not equally *acceptable* for particular audiences. **Exp1** would be obviously *unsuitable* for a grade-school child who would benefit from **Exp2**.[3] Far from *increasing trust* in the assistant's assertions, such an inscrutable response would likely *diminish trust.*

The average child would likely meet the digital assistant's talk of *agouti proteins* and *pheomelanin* with a blank, uncomprehending stare. More generally, the most natural explanation of the unfittingness of **Exp1** for this audience could be its *detours* through *concepts* with which a child *cannot be reasonably be expected to be familiar.* This leads to a very natural demand:

1 An explanation Π should only be presented to an agent α in case α can be expected to grasp every concept appealed to in Π.

Let us say that for a particular notion C—*e.g.*, a *concept* in the case of description logics—its *conceptual complexity* is a collection of the minimal assumptions required of an audience to reliably understand the use of C. *E.g.*, virtually any potential interlocutor can be expected to exercise a command of the concept `Tooth` (given the natural interpretation) while only those with postgraduate training in biology can be expected to have mastery over the concept `Melanocyte`.

[3] Conversely, a PhD holder who is known to study genetics may find **Exp2** too reductive to be useful, but we emphasize the former case.

2 Assigning Measures of Complexity

Intuitively, the concept `Tooth` is far less conceptually complex than `Melanocyte`. In this section, we discuss an approach in the context of the Cyc platform [5] used to assign measures of complexity to terms (and thus of justifications) and to do so efficiently and in bulk. Given an understanding of the background of a user, justifications can be filtered, ordered, or tailored appropriately. Although the implementation was particular to Cyc, the technique applies just as well to other frameworks like OWL 2 [2].

Clearly, some features of conceptual complexity are irreducibly qualitative; successful appeals to concepts involving esoteric surgical techniques presuppose a *familiarity with surgery*. But there exist qualitative assessments that roughly track *e.g.* the *age* or *educational level* of potential interlocutors that can be recovered. One information-rich source that can be mined to determine quantitative evaluations is large natural language corpora like *Project Gutenberg*. A fundamental thesis that guides the manner in which information is drawn from such corpora is the following:

2 Authors typically use concepts in texts only when the intended audience is expected to comprehend them.

This principle is a simple observation about the pragmatic elements of writing. Authors typically are interested in successfully communicating to the intended audience. If an author writes a text that is targeted towards a preschool audience, to appeal to concepts not understood by a typical preschool-age child would go against this interest. There are, of course, exceptions, but *en masse*, we should expect to see a concept C see its first *routine use* across large corpora in those texts targeted towards an appropriate audience.

This brings us to the third thesis:

3 Readability metrics are a good proxy for the conceptual acuity necessary to understand a text, *i.e.*, the intended audience of a text.

Readability metrics—like the Flesch-Kincaid readability test of [4]—are a widely adopted tool in *e.g.* education in order to assess whether a text is appropriate to a grade level. Readability metrics assign a measurement by evaluating factors like *average number of syllables in words* or *average number of words in sentences*. Although these criteria are relatively imprecise, the efficacy of these tools has been confirmed empirically. Given the emphasis on pedagogical applications, the values of such metrics often align with *grade level*, which itself is a sort of proxy for *age* or *intellectual maturity*.

Putting this together begins to reveal the shape of an implementation: A concept C will not frequently enter texts towards audiences not equipped to grasp C. The intended audience of texts can be reliably evaluated by assigning readability values to them. So a reasonable, *initial* proxy for conceptual complexity of C will be the *least readability value* for a class of texts in which lexifications for C (*e.g.* synonyms) begin to appear. If *e.g.* mentions of concept C—or synonymous terms—are more-or-less evenly distributed across texts ranging to preschool to postgraduate, this is a good indicator that it is appropriate to appeal to C in explanations for any audience.

The implementation of this plan in the Cyc system was conducted by the following steps:

1. For a concept C, appeal to WordNet [3] to determine a set Γ_C of synonymous terms
2. For each term in Γ_C, use NLTK Lemmatizer [1] to generate inflectional terms
3. For each text T in Gutenberg, assign a Flesch-Kincaid readability metric n_T
4. Partition the texts according to *grade level* corresponding to n_T
5. For each concept C, determine the least partition at which instances of Γ_C routinely appear

As an illustration, we provide two graphs of the normalized concentrations of Cyc concepts `#$River` and `#$Politician` by approximate grade level. The x-axis is a 2500-point Flesch-Kincaid variant, where the value corresponds to a value one hundred times the corresponding grade level, while the y-axis captures the percent of appearances within a grade level, normalized for number of words.

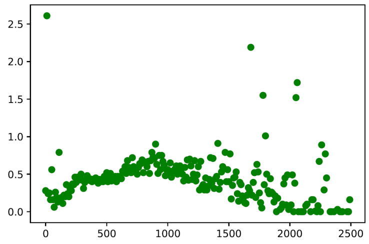

Fig. 1. Distribution on a 2500-point Flesch-Kincaid Analysis of the Cyc term `#$River`

In Figure 1, we see a steep slope up very early on at roughly 400, *i.e.*, fourth grade, suggesting that `#$River` can be safely used in explanations for nearly all potential interlocutors. In contrast, Figure 2 shows a much more subtle slope for the concept `#$Politician`, providing a visual indication of the increase in conceptual complexity over `#$River`.

Conceptual complexity need not be strictly *numeric*, however. Over the course of an education, knowledge becomes *specialized*—although a category theorist and an oncologist have roughly the same *amount* of training, their individual specialties ensure their comprehension of the notions of *epimorphism* and *cytometry*, respectively. In cases in which terms are judged to be at a postgraduate level, the metadata in Gutenberg can be leveraged against *e.g.* an ontology of professions to further refine the assumptions.

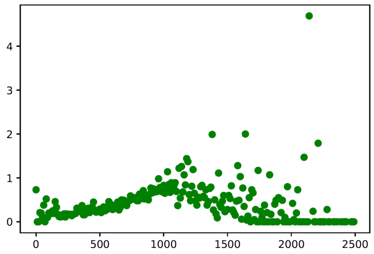

Fig. 2. Distribution on a 2500-point Flesch-Kincaid Analysis of the Cyc term `#$Politician`

3 Some Challenges

Clearly, the procedure described above remains relatively imprecise and several challenges exist. One of the most pressing is the issue of *homographs*; the above technique looks only for particular strings from Γ_C but is indifferent to the *sense* in which the term is being used. In practice, the use of synonymous terms may absorb some of the interference of homographs, but more refined techniques for parsing a corpus in order to control for sense would generate better results. We have also assumed the compositionality of meaning but, as a reviewer has pointed out, this fails to account for non-compositional idioms. Work on representing such idioms in WordNet like [6] could prove useful in incorporating non-compositional terms as well.

A further important point to address is the method of determining numerical value of complexity for a concept C; the plan uses the phrase "routinely appear" but this is clearly underdetermined. Finding a more precise reading of "routinely appearing" is additionally a critical step.

Nevertheless, the above is intended as a sketch of how the steps may be performed. The most important thing is pressing that there is a *need* for such measurements to adequately target explanations and to suggest that there is sufficient information in large corpora to mine these measurements. The fine details can be worked out in the future.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Cambridge (2009)
2. Bock, C., Fokoue, A., Haase, P., Hoekstra, R., Horrocks, I., Ruttenberg, A., Sattler, U., Smith, M.: OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax. W3C, Second edn. (2012), http://www.w3.org/TR/owl2-syntax/
3. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)

4. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Tech. Rep. 8-75, Chief of Naval Technical Training, Naval Air Station Memphis (1975)
5. Lenat, D.B., Guha, R.V.: Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, Reading, MA (1990)
6. Osherson, A., Fellbaum, C.: The representation of idioms in wordnet. In: Global WordNet Conference (2010)
7. Våge, D.I., Fuglei, E., Snipstad, K., Beheima, J., Landsemc, V.M., Klungland, H.: Two cysteine substitutions in the MC1R generate the blue variant of the arctic fox (*Alopex lagopus*) and prevent expression of the white winter coat. Peptides **26**(10), 1814—-1817 (2005)