Choosing a Logic to Represent the Semantics of Natural Language

Adam Pease[1]

Articulate Software, USA, apease@articulatesoftware.com

Abstract. We attempt to answer the question of which kind of logical language should be chosen to represent the semantics of a broad selection of natural language sentences, and how prevalent different kinds of sentences are that require different levels of logical expressiveness. We examine these requirements for representing the semantics of text in logic by studying a sample of several balanced corpora. Our method is to create lists of words and sentential constructs that can easily be assessed in text, which are then mapped to requirements for logics of different expressiveness. We then run an automated analysis on thousands of sentences from two English corpora and manually validate a sample.

1 Introduction

Work in linguistic semantics has often employed logics that are quite expressive, exceeding that of first order logic, typically employing various modal operators [10,4,8]. Work in computer science, particularly in industrial applications, often employs languages of lesser expressiveness, informal approachs such as knowledge graphs [15], or the description logic [1] used in semantic web languages and tools. Implicit in these uses is that the logic employed is sufficient for the task at hand. Tools are often chosen based on some combination of an assessment of prevalence and ease of use. When logics are used to represent a wide domain of knowledge, or used to capture knowledge from a variety of textual sources, it would be beneficial to have quantitative metrics that would indicate the proportion of statements that are expressible in a variety of logics.

There is no system that can automatically convert arbitrary text into an expressive logic, and even human coders will have different interpretations of text, which may, at times, result in statements that require a different logic. However, we can attempt a first exploration in this area, with the hope that this will lead to further studies.

Our approach is to start with looking at particular words that typically require particular logics to capture the semantics of sentences in which they appear. We then collect statistics on those words in different corpora. Lastly, we take a small sample of expressive sentences from a corpus and encode them manually, in order to validate whether the word lists are in fact indicative of the logical constructs we believe are required.

While many researchers such as [9] have shown examples where linguistic semantics requires expressive logics, to date there has not been an automated

quantitative experiment to determine how prevalent such sentences are in large and balanced linguistic corpora. That is what that paper attempts to address.

2 Different Logics

We will only consider a few broad categories or kinds of logics rather than exhausitively considering many specialized logics or variants within these categories. We attempt to show that those categories can be determined from particular words and simple syntactic constructs. We will also assume that we must go beyond a propositional representation. Propositional logic does not allow for use of variables. While it is often possible to create an abstraction of a single sentence that is propositional, once we have a text with multiple sentences, we assume that it will rarely be possible to avoid some need for variables. This will hopefully be clear once we provide example formalizations of some sample sentences.

Logics less expressive than first order logic have only restricted forms of negation and quantification, such as the atomic negation in standard (AL) description logic, so words that lead to these logical features will be the first test for expressiveness.

Beyond standard first-order (barring a special purpose encoding of a first-order modal logic, which we will consider equivalent) are constructs that require a notion of necessity or possibility, as in the S1-S5 families of modal logics [6]. We will attempt to verify this assertion by showing that in our sample, few examples avoid quantification over formulas.

A next level of expressiveness is that of epistemics and authorship expressions, which attribute a text to a particular person.

Note that for each of these logical features, we need not have 100% accuracy in our analysis. It is acceptable to have words missing for each feature, as we aim simply to have a conservative estimate of expressiveness required. If we fail to identify logically expressive sentences that will simply lead to a more conservative assessment. We do however need to be careful about false positives and a manual coding of the identified expressive sentences should help to provide confidence in that regard.

3 Computation and Lexical Semantics

We employ the Suggested Upper Merged Ontology (SUMO) [11][1] and its associated SUO-KIF [12] language due to its large size (roughly 20,000 terms and 80,000 human-authored logical axioms) and expressive representation in a higher order logic. Its use in modern theorem provers allows the theory and extensions to be tested and employed in practical reasoning [2,14,13]. By choosing a more expressive logic we can use a single language and less expressive formulas will simply not take advantage of the full expressiveness of the language. We can also

[1] `http://www.ontologyportal.org`

anchor our terms to an existing defined set of terms in the ontology, and not have to use symbols that have an imagined or intended meaning as opposed to a formal and logically specified one. We can also avoid the impractical alternative of having to define all the symbols used from scratch.

Briefly, in order to interpret the formulas below, SUO-KIF is a prefix notation in a standard Lisp S-expression syntax, where only the seven logical operators ("forall", "exists", "=>", "and", "or", "not", "<=>") plus equality ("=") are reseved words in the language, and all other symbols must be defined in SUMO in terms of those operators. Universal quantification is implicit for unquantified variables. Variables are denoted by a leading '?' sign. In this paper we will highlight terms from SUMO given in the text in `typewriter font`.

4 Word Lists

We rely on the Stanford CoreNLP system [7] to identify *negation*. It is a machine learning based system that was trained on a large set of manually-labeled sentences. To indicate *quantifiers* we select the words "some", "many", "few", "all". We will assess sentences with negation or quantifiers as requiring first order logic.

For *modal* expressions we chose a list of "can", "could", "may", "might", "must", "shall", "should" and "would". Some *other modals*, which appear to be less reliable indicators are "ought", "dare", and "need".

Finally, we have words that indicate statements of knowledge or belief, which we can broadly call *epistemic* operators. These include "know", "think", "learn", "understand", "perceive", "feel", "guess", "recognize", "notice", "want", "wish", "hope", "decide", "expect", "prefer", "remember", "forget", "imagine", and "believe". Statements of *authorship* would require a different operator that takes a formula as an argument, but have the same requirement for logical expressiveness as epistemics. They are "say", and "write".

5 Experiment

We wrote a simple open source program in Java[2] that calls the Stanford CoreNLP system to do sentence segmentation, tokenization, lemmatization and dependency parsing as steps to enable this analysis. Those functions enable Stanford's negation detection component as well as checking for the presence of words in our various word lists. We only count a sentence as being in one particular category even if it has multiple kinds of operators. Execution time is dominated by negation detection because of its upstream reliance on dependency parsing, but is still relatively fast, completing analysis of the Brown corpus [5] in just a few minutes on a modern laptop computer. Our results for the Brown Corpus are shown in Table 1.

[2] `https://github.com/ontologyportal/sigmanlp/blob/master/src/main/java/com/articulate/nlp/corpora/LogicLevel.java`

In addition, since we are not primarily concerned with works of fiction in commercial applications, we chose a comparably sized portion of the newspaper collection from the Corpus of Contemporary American English [3]. Running on just the year 2012 we get comparable results to the Brown corpus tests, which seem to indicate that these logical features are broadly no more or less prevalent in news than in a balanced corpus that includes works of fiction, poetry and spoken text transcripts. These results are shown in Table 2. Note that percentages are rounded and so do not add up to 100%. Note also that the spacing in the examples reflects tokenizing, where tokens such as in "ca n't" are separated by a space.

Type of operator	**count**	%
negation	419	10.00%
epistemic	243	6.00%
modal	666	16.00%
other modal	27	0.66%
quantifier	177	4.30%
authorship	196	4.80%
simple	2304	57.00%
total	4032	

Table 1. Brown Corpus statistics

Type of operator	**count**	%
negation	513	13.00%
epistemic	328	8.60%
modal	369	9.70%
other modal	27	0.71%
quantifier	223	5.90%
authorship	416	11.00%
simple	1897	50.00%
total	3773	

Table 2. COCA 2012 News statistics

6 Experiment Validation

We next selected a random sample of 100 sentences, using Java's `Random` class, that were marked by our automated analysis as not "simple" and attempted to formalize them manually in an expressive logic. Note that we are simply interested in a "upper bound" of how many sentences do not require expressive logics,

so we do not need to perform a manual formalization of any of the sentences in the category of "simple", since if a simple sentence required an expressive formalization that would only decrease the upper bound.

Even in the case of a balanced corpus like the Brown Corpus that includes fiction and poetry, only 57% of sentences are "simple" and without negation, modals, authorship or epistemics. This is also likely to be conservative since we do not consider constructs such as metaphors, some of which can require complex logical representations without the explicit keywords that we have measured.

The first randomly selected sentence (from corpus line 45437) marked as being a statement of "authorship" was

" It 's just a waste of resources , if you ask me , " she said .

We coded the statement in the SUO-KIF logical language, using terms from SUMO. No new terms were needed for this encoding. The interested reader can look at the definitions of these terms by entering them in the online browser[3]. The formalization can be paraphrased as "There is a speaking event, where the agent of the event says that there's an different event that uses a resource, which does not benefit anyone."

There are many possible encodings of this statement, and no doubt many that could be considered "deeper" by explicitly modeling a notion of waste rather than just an absence of benefit, or the implications of politeness or modesty of the phrase "...if you ask me..." But the essential feature relevant to this experiment, which would still be present in other options for formalization, is that the speech has some logical content and stating that content as an explicit logical formula, as opposed to a logically opaque term or proposition, requires logical expressiveness beyond first order logic. The relation in this case is `containsFormula`, which relates a `Physical` thing (which is a class that includes any thing positioned in space and time, and therefore includes `Process`es) and a `Formula`.

```
(exists (?S ?SAY)
  (and
    (instance ?SAY Speaking)
    (agent ?SAY ?SHE)
    (containsFormula ?SAY
      (exists (?IT ?R)
        (and
          (resource ?IT ?R)
          (not
            (exists (?P)
              (benefits ?IT ?P))))))))
```

Statements about propositions are so common that even when one is found there are usually more in the same sentence that aren't signalled by simple keywords, as in

[3] `https://sigma.ontologyportal.org:8443/sigma/Browse.jsp?kb=SUMO&lang=EnglishLanguage&flang=SUO-KIF&term=Speaking`

Ministers were exploring several options to close that gap, but as talks dragged on Monday, no final solution appeared imminent .

which was marked as a negation ("...**no** final solution...") but where *appeared* and *imminent* also state relationships (epistemic and temporal, respectively) to a proposition (that a final agreement will be achieved in the negotiation) that require a higher-order logic.

We now present some of a set of randomly chosen sentences from the COCA 2012 news corpus (file `wlp_news_2012.txt`) the first few along with their formalization in SUO-KIF/SUMO. The line number of the corpus is given and then a keyword for how it was classified on the basis of the different word lists given in the body of the paper. There were two sentences out of our random sample of 100 with critical elisions and two that do not require an expressive logic. The two sentences in our sample of 100 that one could argue have been misclassified are

(corpus line 6504) epistemic: *The testing , to be carried out over the next several weeks , marks a significant expansion of the agency 's probe in Dimock , a tiny crossroads at the center of a national debate over gas drilling and the extraction technique known as hydraulic fracturing , or fracking.*

"...known..." in this case is not an epistemic but just an expression of synonymy. If we create a class of Fracking one could have a simple relation of "`communicationAbout`" that would relate a communication event "...debate..." and a class representing a topic, so it's possible this could be done in a description logic.

(corpus line 6431) neg: *I ca n't even tell you , again , what a relief this is.*

Read literally, "...can't tell..." is a negation but it isn't since later in the sentence the speaker does tell the listener what he or she wants to say, that [it] is a "relief". It's just a politeness construct. If the referent of "it" is a complex statement that would have to be modeled as a formula, then this is HOL. But if it's just an event, then it could be represented in FOL or even DL

Five sentences of the 100 were fully formalized in SUO-KIF/SUMO - corpus line numbers 65640, 9632, 77220, 70553 and 53967.

Note that one additional sentence (corpus line 45437) is formalized in the text above.

(corpus line 65640) neg: *Ministers were exploring several options to close that gap , but as talks dragged on Monday , no final solution appeared imminent .*

```
(exists (?M1 ?M2 ?N ?M)
  (and
    (attribute ?M1 GovernmentPerson)
    (attribute ?M2 GovernmentPerson)
    (not
      (equal ?M1 ?M2))
    (instance ?M Monday)
    (instance ?N Negotiating)
    (during ?N ?M)
    (agent ?N ?M1)
    (agent ?N ?M2)
    (not
      (expects ?M1
        (holdsDuring
          (ImmediateFutureFn
            (WhenFn ?N)
            (exists (?A)
              (and
                (instance ?A Agreement)
                (result ?N ?A)))))))
    (not
      (expects ?M2
        (holdsDuring
          (ImmediateFutureFn
            (WhenFn ?N)
            (exists (?A)
              (and
                (instance ?A Agreement)
                (result ?N ?A)))))))))
```

Note this sentence is already given in the text above but the formalization is given here. Also to note is that we know from the plural 'Ministers' that there is more than one minister involved in the event. But we do not know that there are more than two involved. The logical form created with two different `GovernmentPerson`s in an `agent` relation requires two ministers, but does not entail that there are only two.

7 Conclusion

We reviewed all 98 of the randomly chosen sentences. Two sentences were rejected because the news corpus has some elided phrases, replaced with "@ @ @..." that makes it impossible to provide a complete formalization. For two sentences of the 98, it should be possible for formalize them using only a description logic. We did a "complete" formalization of the first 6 sentences. Of the 90 remaining sentences we reviewed that have required the logic determined by the keyword lists, they usually also require several more advanced logical operators. We have posted these validations on line as an appendix to this paper [4]. We believe that we can reasonably conclude that the statistics given in section 5 are conservative. The results show that roughly half of a the sentences in the test corpus require a logical expressiveness of full first order logic or greater. We hope that this may lead researchers and practitioners to reconsider the choice of less- expressive logics for knowledge representation, or at least be more aware about the limitations they impose on the percentage of human communication that requires greater expressiveness.

References

1. Baader, F., Horrocks, I., Sattler, U.: Description Logics. In: van Harmelen, F., Lifschitz, V., Porter, B. (eds.) Handbook of Knowledge Representation, chap. 3, pp. 135–180. Elsevier (2008), `download/2007/BaHS07a.pdf`
2. Benzmüller, C., Pease, A.: Progress in automating higher-order ontology reasoning. In: Konev, B., Schmidt, R., Schulz, S. (eds.) Workshop on Practical Aspects of Automated Reasoning (PAAR-2010). CEUR Workshop Proceedings, Edinburgh, UK (2010)
3. Davies, M.: The corpus of contemporary american english as the first reliable monitor corpus of english. Literary and linguistic computing **25**(4), 447–464 (2010)
4. Gochet, P., Gribomont, E.P.: Epistemic logic. In: Gabbay, D.M., Woods, J. (eds.) Logic and the Modalities in the Twentieth Century, Handbook of the History of Logic, vol. 7, pp. 99–195. Elsevier (2006)
5. Kucera, H., Francis, W.N.: Computational Analysis of Present-Day American English. Providence: Brown University Press (1967)
6. Lewis, C.I., Langford, C.H.: Symbolic logic,. The Century Co. (1932)
7. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), `http://www.aclweb.org/anthology/P/P14/P14-5010`
8. Mineshima, K., Martínez-Gómez, P., Miyao, Y., Bekki, D.: Higher-order logical inference with compositional semantics. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2055–2061. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1244, `https://www.aclweb.org/anthology/D15-1244`

[4] `https://adampease.org/CLAR20201-appendix.pdf`

9. Montague, R.: The proper treatment of quantification in ordinary English. In: Hintikka, K.J.J., Moravcsic, J., Suppes, P. (eds.) Approaches to Natural Language, pp. 221–242. Reidel, Dordrecht (1973)
10. Moss, L.S., Tiede, H.J.: Applications of modal logic in linguistics. In: Handbook of Modal Logic (2007)
11. Niles, I., Pease, A.: Toward a Standard Upper Ontology. In: Welty, C., Smith, B. (eds.) Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). pp. 2–9 (2001)
12. Pease, A.: SUO-KIF Reference Manual. web document (2009), `https://github.com/ontologyportal/sigmakee/blob/master/suo-kif.pdf`, retrieved 20 June 2020
13. Pease, A.: Arithmetic and inference in a large theory. In: AI in Theorem Proving (2019)
14. Pease, A., Sutcliffe, G., Siegel, N., Trac, S.: Large Theory Reasoning with SUMO at CASC. AI Communications, Special issue on Practical Aspects of Automated Reasoning **23**(2-3), 137–144 (2010)
15. Singhal, A.: Introducing the knowledge graph: Things, not strings (2012), `https://blog.google/products/search/introducing-knowledge-graph-things-not/`