Finding enthymemes in real-world texts: A feasibility study

Olesya Razuvayevskaya* and Simone Teufel

Computer Laboratory, University of Cambridge, United Kingdom
E-mails: or264@cam.ac.uk, sht25@cam.ac.uk

Abstract. Enthymeme reconstruction, i.e. the task of reformulating arguments with missing propositions, is an exciting task at the borderline of text understanding and argument interpretation. However, there is some doubt in the community about the feasibility of this task due to the wide range of possible reformulations that are open to humans, and the subsequent low expected agreement. We therefore believe that research on how to define an objective ground truth for these tasks is necessary before any work on the automatic reconstruction can begin.

Here, we present a feasibility study for the task of finding and expanding enthymemes involving a fortiori arguments in real-world texts, and we show that given a sufficiently strict reformulation of the human annotation task, substantial agreement can be achieved. We split the task into three sub-tasks: 1. deciding whether a candidate text span really represents an enthymematic argument, 2. classifying the type of a fortiori argument concerned and 3. describing the missing premise in natural language. In a case study involving the two authors of this paper as annotators, we test a specific type of a fortiori arguments, the let alone construction, for its suitability for reaching high agreement in all three stages of the task. We also discuss pragmatic effects of let alone and how they relate to argumentation theory.

Keywords: Argumentation, enthymemes, annotation experiment, pragmatics, a fortiori

1. Introduction

According to the Aristotelian definition [6], enthymemes are standard-form syllogisms with one missing proposition. But in the modern usage of this term in argumentation theory, enthymemes are often defined as any type of arguments where one of the principal inferences is missing [2,7,12,45]. These missing inferences often express generally known facts. Consider the following argument:

(1) All reptiles are cold-blooded animals. Therefore, lizards are cold-blooded.

Example (1) is a syllogism where the inference from the first sentence to the second one, lizards are reptiles, is a widely known fact and is not stated explicitly.

Enthymeme resolution is the task of restoring the original meaning of an argument. It often comprises a set of procedures, such as paraphrasing, reformulation, and insertion of generally accepted facts. The computational task of enthymeme reconstruction can be defined as the automatic insertion of missing premises or conclusions in naturally occurring enthymemes. It is an important step of the argument mining pipeline. Mining a collection of documents for arguments is a multistage process of identifying arguments in text, analysing smaller segments within these arguments and understanding relations between those segments. Peldszus and Stede [27] define five hierarchical steps of the task of argument mining.

*Corresponding author. E-mail: or264@cam.ac.uk.

1946-2166/17/$35.00 © 2017 – IOS Press and the authors.

This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License (CC BY-NC 4.0).
mining: argument extraction, segmentation, i.e. identification of minimal argumentative discourse units (ADUs), segment classification, i.e. labeling of ADUs based on their argumentative roles, identification of relations between these segments, and argument completion, i.e. automatic construction of statements from implicit propositions. The majority of currently existing computational techniques are aimed at the first four steps in this hierarchy that are related to the analysis of explicitly specified argumentative information: detection of rhetoric structures [22,41] and argument zones [38–40], extraction of arguments [1,14,32] and detection of argumentative components [9,24]. Our goal is to contribute to the final stage, the problem of detecting and inserting implicitly stated argumentative propositions. It is a new and exciting prospect for text understanding and argumentation theory. The automatic reconstruction of enthymemes would be highly desirable for various applications, such as knowledge mining, i.e. automated acquisition of common knowledge information that can potentially be stored in knowledge bases, argument validity verification, and automatic simplification of written texts for children or adult readers of lower reading age. It would also further our insights into the process of text understanding and argument interpretation, in particular into the question which inference steps are necessary when reading and understanding a text.

However, the literature on argumentation theory is generally pessimistic about the feasibility of an objective standard for enthymeme reconstruction [3,15,17]: the goal has been called a “challenging and subjective task even for human experts, often resulting in a wrong interpretation or exaggeration of the arguments” [34]. Hitchcock [17] mentions two particular problems associated with enthymemes, a) the difficulty of distinguishing enthymemes from deductively valid arguments and from arguments that are to be rejected (“demarcation problem”) and b) the difficulty of evaluating whether the inferred expression was the one that was originally intended.

Due to the known inherent subjectivity of the enthymeme reconstruction task, a ground truth for this problem can in our opinion only be established by means of human annotation. If two or more annotators, working independently, can agree on the existence and identity of the missing premise, we will accept it as a proof that the enthymeme objectively “exists”. This turns the problem of the subjectivity of enthymeme reconstruction into an empirical question, which can be answered simply by measuring agreement. We investigate the feasibility of several sub-tasks associated with this task:

• Agreement on whether the given argument is enthymematic and there exists a missing inference;
• Agreement on which subtype of inference is concerned;
• Agreement on the missing premise explicitly stated in natural language.

The annotation scheme constructed for this study is described in Section 4. In this work, we investigate a specific case of a fortiori arguments that can be interpreted without additional context and with a minimum amount of general knowledge. We discuss the properties of such arguments in Section 3.

The methodology and results of our experimental study are presented in Section 5. During the first step, we verify whether annotators agree on the existence of the missing proposition in a given text piece which is potentially representing an a fortiori argument. Annotators then select the category which best explains which type of proposition is missing, from a fixed set of possible categories. The final step of reformulation of the missing premise is again performed under strict rules. We designed a linguistic template for each category; if a certain category is chosen, annotators produce their reformulation by inserting textual material from the sentence into the associated template. Within the constraints of this procedure, annotators are asked to create a reasonably naturally-sounding sentence.

This methodology attempts to make the task of enthymeme insertion more deterministic by constraining the variety of both reasoning steps and linguistic ways of formulating these steps so that:
(a) Minimal or no general knowledge and external information is needed for argument interpretation;
(b) There exists a unique logical reasoning scheme associated with the enthymeme concerned;
(c) The number of choices for formulating the missing reasoning step is restricted.

2. Properties of enthymemes

2.1. Enthymematic syllogisms

All types of arguments can be expressed in a truncated form, but truncated syllogisms are the most studied types of these [7]. Standard-form syllogisms are convenient for the analysis of enthymemes, because such arguments consist of only three terms and satisfy a set of strict requirements. These requirements uniquely define the form and the order of argumentative propositions and terms [36]. Therefore, given any two syllogistic propositions, the goal of restoring the missing premise or conclusion can be achieved by following a set of deterministic steps. For instance, consider the enthymeme below:

(2) Tweety is a bird. [minor premise]
Therefore she flies. [conclusion]

As can be inferred from the formal definition of a syllogism, this is a categorical syllogism with the minor premise and the conclusion expressed. There exists a general procedure [7] that allows us to conclude that the missing proposition is:

All birds fly. [major premise, universally quantified]

The procedure for restoring the missing proposition is the following:

(a) Define the three terms or categories associated with the syllogism. In Example (2), these terms are: Tweety, bird and fly.
(b) Classify the given terms into major, minor and middle ones. By the definition of a syllogism, the major term is the predicate of the conclusion, and the minor term is its subject. Therefore, Tweety is the minor term and fly is the major one. The remaining term, bird, is called the middle term. The middle term is shared by the major and the minor premises.
(c) Given the major and the minor terms, we can identify the types of the premises. The premise that contains the minor term as its subject is the minor premise. In our example, the first proposition is therefore the minor premise, and we can conclude that the missing proposition is the major one.
(d) Now, in order to construct the missing major premise, we must know that the major premise contains the major term as its predicate and the middle term as its subject. We therefore construct the following premise: Birds fly. Now, to make the syllogism valid, we have to make this premise universally true by means of applying the universal quantifier: All birds fly.

In this way, enthymematic syllogisms allow for a straight-forward and objective definition of the missing proposition: they already physically contain all three terms required for the reconstruction, so that no external information is necessary. This means that the process of reconstructing syllogistic enthymemes can potentially be automated. However, it is almost impossible to find well-formed standard syllogisms in everyday language, as we shall see later.
2.2. Enthymemes in everyday language

Unlike syllogisms, the majority of arguments used in natural language discourses are not formed from a set of structurally pre-defined propositions. Instead, people use a set of reasons to support their claims [29]. These reasons can be non-defeasible, i.e. universally true, or defeasible, i.e. they can be defeated by some new information [25,28]. The propositions left implicit in enthymemes are often defeasible, because the speakers prefer not to state potentially weak reasons explicitly [45]. For example:

(3) Of course he does not know about differential equations. He is a lawyer.

The missing proposition here, all lawyers do not know about differential equations, is clearly not universally true.

Information attacking defeasible reasons is called defeater. There exist two types of defeaters, undercuts and rebuttals. The undercutting defeater attacks the connection existing between the reason and the conclusion, while the rebutting one attacks the conclusion itself. For example, recall the enthymeme mentioned in Example (2). It can be attacked by both the undercut, Penguins are also birds, but they do not fly, and the rebuttal, But tweety is too heavy to fly.

Due to the defeasible nature of argumentative claims used in everyday language, formal notations such as syllogistic logic, cannot be applied for the restoration of naturally occurring enthymemes [10]. For instance, Campbell [4] argues that the majority of Darwin’s claims cannot be expressed through logical arguments. The mechanical approach of reconstructing enthymemes in such a way that they become structurally correct and include the needed assumptions may lead to incorrect interpretations of an argument, i.e. to the insertion of premises not intended by the speaker. In Example (2), All birds fly exemplifies this problem. The statement intended by the speaker might be defeasible: Generally, birds fly. Therefore, our attempt to convert the statement into a well-formed syllogism by means the universal quantifier results in an incorrect statement.

Implicit argumentative propositions are not always defeasible. Some of them represent universally true facts that are too trivial to be stated explicitly [18]. This observation is captured by Grice’s Maxim of Quantity – Do not make your contribution more informative than is required [16]. This pragmatic principle discourages a speaker from making explicit statement of facts that are widely accepted by the audience; in fact, the explicit statement of such facts would sound both trivial and also unnatural to a listener. This is another reason why enthymemes are so frequent in natural language discourses.

Jackson and Jacobs [19] notice that enthymemes also play an important role in maximizing a listener’s agreement, because additional information may contain a subjective point of view and increase the possibility of disagreement. For example:

(4) There is no law against composing music when one has no ideas whatsoever. The music of Wagner, therefore, is perfectly legal. Mark Twain

The missing proposition in this humorous argument, Wagner composed music while having no ideas whatsoever, is subjective, and many people will disagree with it. The author therefore refrained from stating it explicitly.

Detection of hidden and implicitly stated arguments requires a thorough analysis of argumentative structures and strong general or domain-specific knowledge, in contrast to the much easier treatment necessary for enthymematic syllogisms. The possession of domain knowledge becomes particularly critical when enthymemes appear in scientific domains, and their resolution is therefore usually reviewed as an activity for experts in the area [4]. Crick states that enthymemes are frequent in scientific writing,
as a result of the fact that the audience is assumed to consist of experts in the field [8]. Enthymemes make scientific persuasion more engaging and initiate cognitive cooperation between the writer and the audience, hence creating a stronger persuasion effect than if complete arguments had been used, which may be initially surprising. According to Campbell, the effect is achieved by the shared reasoning process between writer and audience, where missing premises are jointly filled in [4]. For this persuasive power, Aristotle called enthymemes “the strongest of rhetorical proofs”.

As can be seen from the reasons that motivate occurrence enthymemes in everyday language, finding the assumptions intended by the speaker without having the formal structure of an argument is an unfeasibly challenging task. It relies on general knowledge and on subjective interpretation of an enthymeme given the unlimited number of choices and humans’ viewpoints [10]. Moreover, it is not clear how many assumptions should be inserted in order to make the argument complete. Therefore, our solution relies on a fixed argumentation strategy, thereby carefully restraining the theoretically possibly unlimited choice. As a result, we do not require a fully mechanical procedure for enthymeme reconstruction.

3. **A fortiori** arguments

In this study, we concentrate of *a fortiori* arguments (first mentioned in Aristotle’s Rhetoric [6]). These arguments can be classified in two main groups: positive and negative.

The main principle behind negative *a fortiori* arguments is that there is a case where a certain quality is more likely to exist than in some other, less likely case. If we now assert that even in the more likely case, the quality does not exist, the logical conclusion is that the quality certainly cannot exist in the less likely case. The *a fortiori* argument therefore serves as a refutation of the less likely case. This type of *a fortiori* argument is called *a minore ad maius*, i.e., from smaller to bigger.

An equivalent case called *a maiore ad minus* exists for positive *a fortiori* reasoning. This is based on the inference about the existence of the quality for the more likely case, based on the fact that even the less likely case has this quality.

3.1. Kienpointner’s [20] scheme and extension

Kienpointner [20] designed the following argumentation scheme for *a minore ad maius* arguments:

- If even X does not have property P and it is a less likely case that Y has property P than that X has property P, then Y does not have P
- (Even) X does not have P
- (Therefore) Y does not have P

We formulate the equivalent scheme for *a maiore ad minus*:

- If even X has property P and it is a more likely case that Y has property P than that X has property P, then Y has P
- (Even) X has P
- (Therefore) Y has P

Our observation is that while this scheme is correct and explanatory, it does not express additional implicit scaling relation holding between X and Y. Consider the following example:

(5) **If even the house could not accommodate them, the room definitely will not.**
Kienpointer’s scheme only allows expressing:

If even the house does not have a property of accommodating them, and it is a less likely case that the room has a property of accommodating them than the house has the property of accommodating them, then the room does not have a property of accommodating them.

As can be seen, the suggested insertion does not explain the key scaling relation between X and Y:

The size of the room is smaller than the size of the house.

To address this problem, in this study, we extend Kienpointer’s argumentation scheme with a fourth proposition that describes a concrete relations holding between X and Y:

X and Y stand in Relation R.

The relation R between X and Y is almost always implicit, which makes a fortiori arguments ideal cases for our investigation of the feasibility of reaching agreement in the enthymeme completion task.

3.2. Nedum and let alone sentences

There exist many linguistic constructions for expressing a fortiori arguments in natural language: of course, not...even, if...then, etc.

(6) My child can already write. Of course he knows the alphabet.
(7) I could not run even 3 miles. I definitely will not be able to run this marathon.
(8) If John could solve this task, then you for sure will.

For our study, we investigate one of the most typical a fortiori constructions: let alone sentences. Let alone is a convenient a fortiori construction, because all the necessary components are usually present in a single sentence, and the amount of argument-irrelevant information in that sentence is minimal. Consider Example (5) paraphrased for let alone construction:

(9) Even the house could not accommodate them, let alone the room.

Let alone is a nedum – a connective particle that signalizes that the second clause of the construction is more relevant to the context and can be inferred from the first clause [33]. For example:

(10) This task is difficult for an adult, let alone a child.
(11) The baby cannot sit yet, let alone walk.

According to Fillmore et al. [13], let alone can be syntactically treated as a coordinating conjunction, where the phrases linked by let alone display parallelism with respect to grammatical functions; e.g. direct object (adult and child) or verbal complement to auxiliary (sit and walk). In the terminology of Toosarvandani [42], the stressed elements in the first and second part of the sentence are called correlate and remnant respectively.

Let alone can be used to signal both positive and negative a fortiori arguments, representing a minore ad maius and a maiore ad minus arguments respectively. Sawada [31] distinguishes the following three main cases of let alone arguments: explicit negative (Example (12)), implicit negative (Example (13)),
Table 1

 Annotation scheme for let alone scales

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Part of</td>
</tr>
<tr>
<td>2</td>
<td>Smaller than</td>
</tr>
<tr>
<td>3</td>
<td>Precondition for</td>
</tr>
<tr>
<td>4</td>
<td>Other lexical entailment</td>
</tr>
<tr>
<td>5</td>
<td>Earlier date</td>
</tr>
<tr>
<td>6</td>
<td>Additional constraint</td>
</tr>
<tr>
<td>7</td>
<td>Additional referents</td>
</tr>
<tr>
<td>8</td>
<td>Cumulative/independent</td>
</tr>
<tr>
<td>9</td>
<td>More extreme case than</td>
</tr>
<tr>
<td>10</td>
<td>Easier than</td>
</tr>
<tr>
<td>11</td>
<td>Less likely than</td>
</tr>
</tbody>
</table>

and explicit positive (Example (14)):

(12) He cannot solve simple, let alone starred tasks.
(13) The task is too difficult to be understood, let alone solved.
(14) He could find the exact answer, let alone an approximate solution.

Toosarvandani [42] observes that some form of scaling information is always present between the remnant and the correlate, making the correlate in some respect more likely for negative sentences and less likely for positive ones. The hearer can pragmatically infer, based on common knowledge [16,35], that these stressed elements are ordered based on some scale. This hidden scalar reasoning is highly relevant to our task of finding arguments with hidden premises, and is captured in the extension to Kienpointner’s scheme presented in Section 3.1 above. We argue that let alone sentences are by definition enthymemes, and that the missing scalar relation, called relation R above, is an important part of the missing proposition necessary for a conclusion to be accepted. The conclusion would therefore always be a statement about the remnant, which is entailed from both the parallel statement about the correlate and the inserted scalar relation.

4. Annotation scheme

We categorised the scaling relations based on a development corpus of about 250 cases of let alone sentences, resulting in the scheme in Table 1.¹

The first 8 categories in our scheme were motivated by the most frequent let alone cases observed in the development corpus. These categories are neither mutually exclusive nor are they at the same level, but they are in our opinion the ones that characterize the various phenomena best. For example, any pair of remnant and correlate standing in relation part of will by default also satisfy the smaller than relation. The general annotation rule therefore was to select the highest-priority category that describes the situation. We define the priority by specificity and order the categories from more specific to less specific. As part of is more specific than smaller than, the annotators would have to give it preference.

¹Agreement reported in this paper is measured on a separate test corpus.
For the types of scalar relations that rely on very subjective inferences or the ones that were not observed in the development set but can occur during the testing time, we have identified three fallback categories 9–11 that are more generic than the main ones.2

We will now describe the categories.

4.1. Category definition

1. Part of: The relationship between the referents in remnant and correlate is that of superset. For example:

(15)  This does not apply to Germany, let alone Europe.

The scaling relation $R$ here is: $\text{Germany}$ is a part of $\text{Europe}$.

This relation has a number of requirements:

(1) The statement holds for the holonym (whole) only if it holds for all of its meronyms (components).
(2) The meronym does not have any property that would make it an exceptional case in relation to the rest of the cases (as this is covered in category Additional constraint below). For example, if we know that Germany is much less likely to accept a particular law than the rest of the Europe, we could make a different argument from superset to the subset:

(16)  No country in Europe will accept this law, let alone Germany.

This case is not covered here, but in the category called Additional constraint.

2. Smaller than: The scale concerned involves cardinalities or standard measurements which connect the remnant and the correlate. For example:

(17)  You wouldn’t make to New York, let alone the West Coast.

(18)  There is no space in the fridge for a bottle of water, let alone a saucepan.

The scaling relations here would be:

the distance to New York is smaller than the distance to the West Coast

a bottle of water is smaller than a saucepan

Such arguments compare measurements of the correlate and remnant’s properties (weight, age, volume, etc). In an Example (18), the scalar comparison is between the volumes of a bottle of water and a saucepan:

the volume of a bottle of water is smaller than the volume of a saucepan

However, in order to make the annotation task more straightforward, we omit the obligatory mention of the property and accept the direct scalar comparison between the remnant and the correlate.

3. Precondition: The action in the remnant necessarily requires the action in the correlate having taken place earlier. For example:

(19)  Your talent isn’t enough to participate, let alone win.

2Our choice of three categories rather than one was motivated by our wish to create naturally-sounding template sentences. However, in future work, we will choose only one fallback category, at the cost of somewhat less natural template sentences.
The scaling relation here would be: *participating* is a precondition for *winning*.

Fellbaum [11] presents four types of entailment relations between verbs; our *precondition* corresponds to her *backward presupposition* (*winning* presupposes *participating*).

4. **Other lexical entailment:** This relation covers the larger class of actions where the proposition expressed in the remnant lexically entails the proposition expressed in the correlate. There can be temporal overlap between the two propositions, unlike in category 3, where the proposition in the correlate has to be completed before the proposition in remnant begins.³ For example:

> (20) *He doesn’t even sleep, let alone snore.*

The scaling relation here would be: *snoring* entails *sleeping*. *Snoring* can only happen during *sleeping*. Additionally, *snoring* and *sleeping* can occur simultaneously for some of the time, but it is not necessary that *snoring* entirely overlaps with *sleeping*; there can be times of *sleeping* without *snoring*.

5. **Earlier date:** This is a special case of *Smaller than* where the property is time: in the case of irreversible events, if a state first becomes true at time $t_0$, it by definition does not hold at any earlier time $t < t_0$. This relation relies on the directionality of this particular scale (time moves only in one direction). For example:

> (21) *They didn’t have electricity in 1923, let alone 1909.*

The scaling relation here would be: *1909* is an earlier date than *1923*.

6. **Additional constraint:** In his category, a number of semantic predicates (constraints) applied to the correlate are compared with the same set of predicates plus additional predicates in the remnant. For example:

> (22) *You don’t know what a middle-aged person feels like, let alone a middle-aged prince.*

Here, a *middle-aged prince* is a double-constraint on a person, which filters out some middle-aged non-princes.

7. **Additional referents:** This category covers cases where the set of referents in the remnant logically includes the referent from the correlate. For example:

> (23) *The company does not even insure their employees, let alone their families.*

It is hard to imagine a company that would insure the family of an employee without also insuring the employee himself. The intended meaning of the remnant must therefore be logically interpreted as “the employee and his family”.

8. **Cumulative/independent:** Here, the remnant of *let alone* is not directly comparable to the correlate, but becomes interpretable if we read it as an additive constraint. For example:

> (24) *This Easter-egg packaging does not even protect its contents, let alone have anything to do with Easter.*

³This is analogous to the union of Fellbaum’s categories *co-extensiveness* and *proper temporal inclusion*. *Co-extensiveness* of events holds in one direction for the entire time period (*moving* always happens whenever *walking* is taking place). In contrast, *proper temporal inclusion* may hold between two events that happen to be performed simultaneously, but it is not necessary that both overlap for the entire duration of one of the actions. For instance, *snoring* is possible during some *sleeping* interval. The distinction between these classes is known to be difficult even for humans [11,43], and of no consequence to our task, so we collapsed them.
Table 2

Linguistic templates

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
<th>Linguistic template</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Part of X is a part of Y</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Smaller than X is smaller than Y</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Precondition for X is a precondition for Y</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Other lexical entailment Y lexically entails X</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Earlier date Y is earlier than X</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Additional constraint Y poses additional constraint on X</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Additional referents Y in addition contains X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Cumulative/independent X and Y are cumulatively more important than X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>More extreme case than Y is a more extreme case than X</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Easier than X is easier than Y</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Less likely than Y is less likely than X</td>
<td></td>
</tr>
</tbody>
</table>

Here, both the feature of protecting the contents and of being decorated with Easter motives are important for Easter egg packaging. The intended meaning that speaker put into this argument is:

*This Easter-egg packaging does not even protect its contents. In addition, it does not have anything to do with Easter.*

Hence, the combination of the two qualities of the Easter-egg packaging is more important than only the quality of protecting its contents. One possible test for whether the relation is independent/cumulative is to try to substitute *let alone* with *not mentioning*. If the meaning does not change, then the remnant and the correlate are likely to be in the independent/cumulative relation.

The three categories *Easier than*, *Less likely than*, and *More extreme case than* are fallback options, which are only to be used if none of the more specific categories applies. For example:

(25) *They refused to refer to Kursk, let alone Moscow.*

(26) *I could not solve the first, let alone the last tasks.*

(27) *I have not even seen Mary, let alone Rose.*

4.2. Linguistic templates

There is a connection between the scale category introduced in the previous section and the linguistic form of premise that can be constructed from that class, as can be seen from Table 2.

Each template represents one of many possible verbalizations of the argument associated with each scaling relation. Annotators were asked to extract the appropriate parts from the given sentence and insert them into the template of their chosen category. For instance, for Examples (15)–(27) they were expected to write the sentences in Table 3. The material directly excised from the sentences is written in boldface.

As can be seen from these examples, some of the constructed facts are always true, such as: *Snoring lexically entails sleeping*, *Germany is a part of Europe* or *participating is a precondition for winning*. However, some of the statements are only true in the given context. For example, *The distance to New York* is only smaller than *the distance to the West Coast* for a particular current location of the speaker. Nevertheless, such statements can be good candidates for knowledge mining about generally accepted facts.
Table 3
Filled linguistic templates for categories from Table 1

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
<th>Verbalized statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Part of</td>
<td>Germany is a part of Europe.</td>
</tr>
<tr>
<td>2</td>
<td>Smaller than</td>
<td>The distance to New York is smaller than the distance to the West Coast.</td>
</tr>
<tr>
<td>3</td>
<td>Precondition for</td>
<td>Participating is a precondition for winning.</td>
</tr>
<tr>
<td>4</td>
<td>Other lexical entailment</td>
<td>Snoring lexically entails sleeping.</td>
</tr>
<tr>
<td>5</td>
<td>Earlier date</td>
<td>1909 is earlier than 1923.</td>
</tr>
<tr>
<td>6</td>
<td>Additional constraint</td>
<td>Being a middle aged prince poses additional constraints on being a middle-aged person.</td>
</tr>
<tr>
<td>7</td>
<td>Additional referents</td>
<td>The families of employees in addition contain employees.</td>
</tr>
<tr>
<td>8</td>
<td>Cumulative/independent</td>
<td>Protecting the contents and having relation to Easter are cumulatively more important than protecting the contents.</td>
</tr>
<tr>
<td>9</td>
<td>More extreme case than</td>
<td>Moscow is a more extreme case than Kursk.</td>
</tr>
<tr>
<td>10</td>
<td>Easier than</td>
<td>The first task is easier than the last task.</td>
</tr>
<tr>
<td>11</td>
<td>Less likely than</td>
<td>Seeing Rose there is less likely than seeing Mary there.</td>
</tr>
</tbody>
</table>

5. Annotation experiment

100 random let alone sentences were extracted from the British National Corpus (BNC). The annotation experiment consisted of three sub-tasks:

(a) Classification of the given sentences into enthymemes or non-enthymemes;
(b) Classification of the relation type holding between the remnant and the correlate (as described in Section 4 above);
(c) Filling the corresponding linguistic template with linguistic material (as described in Section 4 above).

5.1. Enthymeme classification task and results

Two annotators (the authors of this paper) made a binary decision when faced with a let alone sentence; they had to decide whether there was enough context to interpret the sentence and to restore the missing statement. They additionally crossed out any information that was not relevant (or actively distracting) to the argument expressed in the let alone sentence.

The number of positive choices was high (183 out of 200). The inter-rater agreement was calculated by means of Cohen’s kappa statistics [5], to be $K = 0.729$ ($N = 100, k = 2, n = 2$). We use Cohen’s standard formula:

$$k = \frac{p_o - p_e}{1 - p_e}$$

$p_o$ is the relative observed agreement among raters, and $p_e$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. The numerator represents the difference between the observed probability of success and the probability of success under the assumption of an extremely bad case. Independence implies that pair of raters agree about as often as two pairs of people who effectively flip coins to make their ratings. Kappa’s possible values are constrained to the interval $[-1; 1]$. $K < 0$ means that the observed agreement is less than the one expected by chance, $K = 0$ means that agreement is not different from chance, and $K = 1$ means perfect agreement. However, just obtaining kappa significantly greater than zero is not
There were only 2 out of 183 cases where either annotator specified enthymeme-irrelevant information, namely:

(28) According to Rex, we never had a chance of negotiating the bend, let alone stopping, given how fast we were travelling.

(29) ‘But the king is not yet married, let alone having a son,’ objected Anne—but then her face brightened.

5.2. Enthymeme classification task

The next stage of the experiment is performed only on those let alone sentences that were positively classified by both annotators as enthymematic (90 cases).

The annotators classified the let alone sentences according to the annotation scheme from Table 1 in order to describe the scale that applies for the a fortiori argument.

The agreement between annotators was $K = 0.75$ ($k = 2, n = 2, N = 90$). This indicates that annotators were overall well able to distinguish relation categories (scales).

The class distribution per category, as presented in Table 4, is skewed. The most frequently chosen category was the “precondition for” relation – 61 (33.9% of cases). The “additional constraint” and “additional referent” classes were also frequent (12.7% and 10% respectively). The fact that the fallback categories were chosen relatively frequently suggests that there are cases where the context is not enough to disambiguate the scale.

The small sample size prevents us from analyzing problematic categories systematically, but we noticed that earlier date and precondition for seem to be relatively harder for the annotators to distinguish reliably. Despite the very distinctive nature of the “cumulative/independent” category, the annotators frequently disagreed on this relation type. We believe that most of the disagreements for the fallback categories came from the fact that one of the annotators preferred these generic relations to more specific ones. For example:

### Table 4

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Number of choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part of</td>
<td>8 (4.4%)</td>
</tr>
<tr>
<td>Smaller than</td>
<td>9 (5%)</td>
</tr>
<tr>
<td>Precondition for</td>
<td>61 (33.9%)</td>
</tr>
<tr>
<td>Other lexical entailment with</td>
<td>11 (6%)</td>
</tr>
<tr>
<td>Earlier date</td>
<td>5 (2.7%)</td>
</tr>
<tr>
<td>Additional constraint on</td>
<td>23 (12.7%)</td>
</tr>
<tr>
<td>Additional referent</td>
<td>18 (10%)</td>
</tr>
<tr>
<td>Cumulative/independent</td>
<td>9 (5%)</td>
</tr>
<tr>
<td>Less likely than</td>
<td>9 (5%)</td>
</tr>
<tr>
<td>More extreme case than</td>
<td>12 (6.7%)</td>
</tr>
<tr>
<td>Easier than</td>
<td>15 (8.3%)</td>
</tr>
</tbody>
</table>

sufficient to evaluate the quality of the agreement. Various schemes to assess kappa’s significance have been proposed [30, 44]. According to the strictest scale proposed by Krippendorff, our annotation was “marginally reliable” and well within the range of what is accepted in Computational Linguistics as good agreement for pragmatics and discourse-level annotation [21].
They couldn’t organise a strike, let alone a revolution.

While one of the annotators classified this relation as a strike is smaller than a revolution, the second participant preferred the easier than category (a strike is easier than a revolution).

While such cases are in the nature of annotating subjective interpretative phenomena and do occur, the relatively high agreement we observe means that overall they do not occur very often.

5.3. Enthymeme reformulation task

The annotators next composed a new sentence expressing the missing premise, from textual material extracted from the given sentence, together with the template for the chosen category, as described in Section 4. The 71 cases were considered where the annotators agreed on the chosen scale category.

We report the percentage of identical or near-identical statements. Judgments of what is identical were made by the first author, but they were very strict, allowing variation only with respect to near-synonyms or the addition or omission of non-relevant information.

69 premises (97%) were judged to be identical or near-identical.

Only two disagreements occurred, the first of which was:

Few adherents to the new classical macroeconomics trouble even to question it, let alone provide an analytical basis to justify it.

While both annotators identified that “questioning” is a precondition for “providing an analytical basis to justify”, they specified partially different propositions:

(a) Questioning something is a precondition for providing an analytical basis to justify something.
(b) Questioning classical macroeconomics is a precondition for providing an analytical basis to justify classical macroeconomics.

The second enthymeme on which the annotators disagreed was:

It’s not exactly as if Wimbledon are an English force – let alone European.

Here, both of the premises were still nearly the same, but the annotators disagreed on the ellipsis resolution:

(a) For Wimbledon, to be an English force is easier than to be a European force.
(b) To be an English force is easier than to be a European force.

Overall, two disagreement cases out of 73 is a surprisingly low number, given that we demanded linguistic descriptions of quite complex pragmatic effects, and given that the annotators, while knowing the scheme and the ideas, were still working entirely independently on a task that was hitherto deemed too unrestricted to be done at all. The pilot experiment in our opinion shows that applying the right kind of restriction can result in an objective definition of truth, even in seemingly open-ended situations.

6. Discussion and conclusions

In this work, we presented our first study on whether it is possible for humans to reliably find and reconstruct enthymemes in unrestricted texts. The study is limited in size and uses the two authors as annotators, but we nevertheless hope to have demonstrated that the a fortiori reasoning exemplified by the let alone construction is well-defined and can be intuitively described using human-directed rules.
This is encouraging, particularly as the community is generally pessimistic about the feasibility of the enthymeme reinsertion task. We are also satisfied with the result from a practical viewpoint because a reliable gold standard is required as evaluation material for our next stages of automatic enthymeme recognition.

For the case of let alone, we have observed high inter-annotator agreement on the decision of whether the text units automatically found are enthymemes. We further investigated whether the annotators also agree on the exact type of pragmatic relation that holds between two propositions. The results show that although there exist some types of relations that are not easily distinguishable from each other, and some that do lack context for interpretation, the annotators are overall able to achieve significant agreement on this task. In a third task of premise reformulation, the agreement was again near-perfect. This result suggests that let alone constructions are ideal candidates for the task of detecting naturally occurring enthymemes in texts and reconstructing the missing propositions. It also means that it is possible to locate all the information needed to reconstruct an enthymeme in the context of the sentence itself.

Let alone sentences are convenient for the analysis of enthymemes because the information we need for interpreting the a fortiori argument is very localized, i.e., both the conclusion and the premise are usually located in one sentence. This reduces the amount of argument-irrelevant information. However, in the general case we would expect that other linguistic structures, such as arguments expressed with because, of course, therefore, occupy much larger text spans, resulting in a larger amount of argument-irrelevant information. For these more complicated constructions, human agreement is unlikely to be as good.

7. Future work

In future work, we plan to perform this task on a larger scale, with independent annotators and on a larger dataset. When doing so, we plan to critically re-consider the relation types in the light of the annotation results, despite the fact that the annotation scheme presented in this study demonstrated a substantial inter-annotator agreement regarding the relations holding between the remnant and the correlate.

As has been mentioned, scalar relations usually deal with the degrees to which the remnant and the correlate have a particular property. Although we omitted the notion of a property in this annotation scheme, we consider strengthening its role and introducing it into the template, because it would be much more explanatory of the factors affecting the interpretation of arguments. Another motivation for asking for explicit determination of the property is that it can be positively or negatively correlated with the context, and this can again add explanatory power when distinguishing positive a fortiori cases from negative ones.

The fact that a fortiori arguments can comprise both positive and negative pragmatic reasoning should be reflected in the annotation instructions. This may result in constructing separate templates for positive and negative inferences.

It is also essential to re-consider all types of relations that have demonstrated low agreement rate in a way that will make them more intuitive for future annotators. For instance, the part of relation was often confused with the additional constraints relation. These two categories can be merged into one in the future, considering that they both describe a relation from general to particular.

Finally, the three fallback categories are very similar to each other. Our original reason for differentiating them was to ensure that the respective templates sounded maximally natural. But weighted against
the redundancy contained in three functionally identical categories, we will in the future opt for a single fallback category.

We could also improve the experiment in other ways. For instance, in the current experiment, when judging the paraphrases of missed premises, there was almost no subjectivity involved, as most variations were objectively the same or very close to identical. However, in the future, with more challenging discourse markers, this may not always be the case. We therefore intend to outsource the decision of whether two paraphrases are identical or near-identical to unrelated humans, e.g., on Mechanical Turk [26]. This would eliminate any bias that might be present in the first author.

As a next step in our research, we plan to expand the range of discourse markers used for a fortiori constructions beyond let alone, but we suspect that not all possible linguistic constructions are plausible candidates for enthymeme resolution. Other cases may not be as context-minimal as let alone sentences, i.e. the reason supporting the claim may often be located not in the same sentence. This may add the amount of argument-irrelevant information and reduce the objectivity of enthymeme interpretation. Moreover, other discourse markers may be less a fortiori-specific, i.e., they may be used to express different logical patterns. Consider the following two cases:

A must be true, because A has never observed to be false
A must be true, because X has told so

While the same because discourse marker is used in both cases, these enthymematic arguments are based on entirely different reasoning types. The first argument is based on the lack of evidence about any controversial case, whereas the second argument is based on the reliability of the claim’s source. These arguments also correspond to different argument schemes, argument from expert opinion and argument from ignorance; therefore, different reconstruction techniques must be applied to these enthymemes. We predict that in order to increase the objectivity of the enthymeme resolution task, the linguistic constructions must uniquely identify their associated argument schemes. It is therefore important to investigate a wide range of other linguistic structures in order to find other a fortiori behaviour similar to nedum.

Human annotation is only the first step towards automatic reconstruction of enthymemes. In particular, automatic resolution of gapping is a challenging task. Syntactic information is often not enough for intelligent reconstruction, and the semantic information is not available automatically [23,37]. We will investigate machine learning from syntactic structure and semantic information in the future.

Another insight that this study suggests is the fact that pragmatic information, which is not traditionally used in argument mining, plays an important role in the resolution of this type of argument. We predict that pragmatic features, if correctly modeled, should in the future beneficially influence the machine learning of argument structure.

References


[research-article] p. 15/17


