A Parallel Corpus of Natural Language and Isabelle Artefacts

Anthony Bordg, Yiannos A. Stathopoulos, and Lawrence C. Paulson

Department of Computer Science and Technology, University of Cambridge, UK

[apdb3,yas23,lp15]@cam.ac.uk

Parallel corpora are key resources for machine translation in natural language processing (NLP). A parallel corpus maps textual scripts in one language (e.g., French) to their equivalents in another language (e.g., English). The language-paired scripts in a parallel corpus are data points used to train language models that learn how to translate text from one language to the other.

Recently, the theorem proving community explored autoformalisation – the task of generating formal proofs that can be recognised by a theorem prover from their counterparts expressed in informal natural language – as an instance of machine translation [1, 2]. Large transformer models, such as Codex [3], have demonstrated that machines can learn to generate code from natural language text through the use of large (parallel) corpora.

We introduce the Isabelle Parallel Corpus (IPC) of natural language and Isabelle/HOL proofs. Natural language proofs in our corpus are expressed using sentences in the natural language of mathematics, with mathematical expressions transcribed using \LaTeX. The aforementioned textual proofs have been extracted from textbooks, International Olympiad of Mathematics solution sheets and other real-world mathematics resources.

In this presentation we will describe our multi-stage approach for constructing our corpus, showcase our annotation tools and discuss the challenges involved in designing the annotation scheme of a parallel corpus linking natural language to formal proofs.

We developed an annotation tool that allows us to (a) record information about artefacts in the corpus, (b) collect parallel natural language and Isabelle/Isar scripts and (c) implement the annotation scheme for the IPC. Our tool is built on top of a special instance of the SErAPIS search engine for Isabelle and supports multi-user annotation.

In the first phase of building our corpus we have sourced over 500 Isabelle artefacts, including theorems, definitions, lemmata and proof scripts. For each artefact we record information that includes a statement of each artefact in the natural language of mathematics typeset in \LaTeX, a \LaTeX citation to the source material (textbooks, journal etc), the page and number (e.g., Theorem 4.1) as they appear in the source material. The second phase, which is ongoing, involves attaching informal and formal Isabelle/Isar proofs to the recorded statements. At the time of writing, we have paired Isabelle/Isar proofs with corresponding informal proofs for 18 statements.

The consensus in NLP is that machine translation models benefit from word and sentence alignments [4, 5]. A sentence alignment for two parallel text scripts in different languages is a pairing that links sentences in one language to sentences in the other language. Similarly, a word alignment links tokens from a script in one language to the tokens of its equivalent script in the other language. The parallel corpus designers are responsible for including annotations
for sentence and word alignments if this information is required by the intended use of the corpus.

However, without answering questions like “Does every sentence in a natural language proof correspond to a statement in Isabelle/Isar?” and “Can one Isabelle/Isar statement account for multiple natural language sentences?”, the nature of sentence and word alignments for a parallel corpus like the IPC is unclear.

Therefore, the first challenge in designing the IPC is to identify the annotation requirements of the corpus for aligning natural language sentences to Isabelle/Isar statements. We conducted a pilot study to determine the requirements of such an annotation and made some observations, including:

1. there are sentences in the natural language that do not correspond to any statement in Isabelle/Isar and vice-versa and this occurs, for example, when the source text and the Isabelle formalisation assume different prerequisites;
2. there is a many-to-many mapping (i.e. not a perfect one-to-one correspondence) between facts within the textual proof of a statement and facts within the corresponding Isabelle/Isar proof script;
3. it sometimes happens that results embedded in Isabelle proofs are not factorised as lemmata, which could be possibly useful results on their own, but this phenomenon does not occur in natural language proofs since one can always refer to a result even if it is not explicitly factorised;
4. both textual and formal proofs may import dependencies in their argumentation. Dependencies in Isabelle/Isar proofs may span multiple theory files.

Our observations give rise to the second challenge in designing IPC: how should dependencies in parallel textual and Isabelle/Isar proofs be incorporated in the corpus? One solution would be to integrate dependencies in the corpus and include data about the reference graph between artefacts [6].

Unlike general-purpose natural language, the language of mathematics follows its own conventions and is interspersed with mathematical expressions [7]. Similarly, proofs in the Isabelle/Isar language are structured and include statements with terms representing assumptions and symbolic reasoning. Therefore, the third challenge is designing a suitable annotation scheme for (a) representing Isabelle/Isar terms and mathematical expressions in textual proofs and (b) establishing an alignment between them. Attractive solutions come from Mathematical Knowledge Management (MKM) and code understanding and generation. For instance, mathematical expressions and terms can be encoded using Presentation and Content MathML [8]. Furthermore, we can overlay our corpus with token type and other information, such as identifier tagging (IT), that will allow researchers to implement masked span/identifier prediction [9] and skip-tree training [10] in models trained with our corpus. We envisage that the third phase of our process will address these challenges and introduce sentence and token alignments to the IPC.

The IPC will be made public on GitHub prior to our presentation in the hope that phase 2 material (data linking natural language proofs to their Isabelle/Isar counterparts) will be useful to researchers in machine learning for theorem proving. We intend to continuously update the corpus (e.g. with sentence and token alignments in phase 3) and this strategy reflects our vision that the IPC is a living corpus with standardised releases to facilitate comparative analysis of machine learning models. We also intend to open our annotation tools to the wider community and we invite all Isabelle users to join in the annotation effort to continuously expand the IPC.
References


