TheoremKB: Towards a Knowledge Base of Mathematical Results

Pierre Senellart

13 December 2019

SinFra Symposium 2019 on Artificial Intelligence

Outline

Brief presentation of PRAIRIE

TheoremKB

PR[AI]RIE

PaRis Artificial Intelligence Research InstitutE

- A “3IA” institute, created on September 1st, 2019
- Academic partners

  cnrs · Inria inventeurs du monde numérique · Institut Pasteur · PSL UNIVERSITÉ PARIS · Université de Paris

- Industry partners

  amazon · DeepMind · engie · facebook · faurecia · GE Healthcare · Google · IDEMIA

  Microsoft · LABS NAVER LABS · NOKIA Bell Labs · Pfizer · PSA · suez · Uber · Valeo · VERTEX

- Transverse project, federating several institutions within Paris
- `https://prairie-institute.fr/`

PR[AI]RIE

PaRis Artificial Intelligence Research InstitutE

- 45 chairs, focusing on:

  Core AI research (scalable, reliable, explainable AI) : autonomous and multi-agent systems; computer vision; data science; machine learning and optimization; natural language processing; networked data management; robotics

  Interdisciplinary research: interfaces with biology, cognitive science, medicine, digital humanities, medicine, physics, social sciences
- Education program, mostly at the M.Sc. and Ph.D. (for specialists, in maths and computer science; for non-specialists from other fields)
- Collaboration with industrial partners, outreach

Outline

Brief presentation of PRAIRIE

TheoremKB

Scope of the project

- Mathematical sciences: mathematics, theoretical computer science, mathematical physics...
- Scientific knowledge in these fields: collection of PDF articles, consisting in particular of:
  - Definitions of concepts, introduction of notation
  - Results (theorems, propositions, lemmas, etc.)
  - Proofs, with references to papers used
  - References to other papers for definitions, results, etc.
- Project in its infancy: collaborators welcome!

Access to scientific literature today…

- Academic search engines (Google Scholar, Microsoft Academic, etc.)
- Bibliographical databases (MathSciNet, DBLP, Scopus, etc.)
- Search by:
    - keywords
    - basic metadata (authors, venues, dates)
    - citation links
- Almost no semantic knowledge

… and things we cannot do

- What variants of the vertex cover problem have been shown to be polynomial-time?
- Are there lemmas in this paper that are actually unused?
- How many published results depend on a given theorem, and what would be the impact if this theorem were discovered to be wrong?

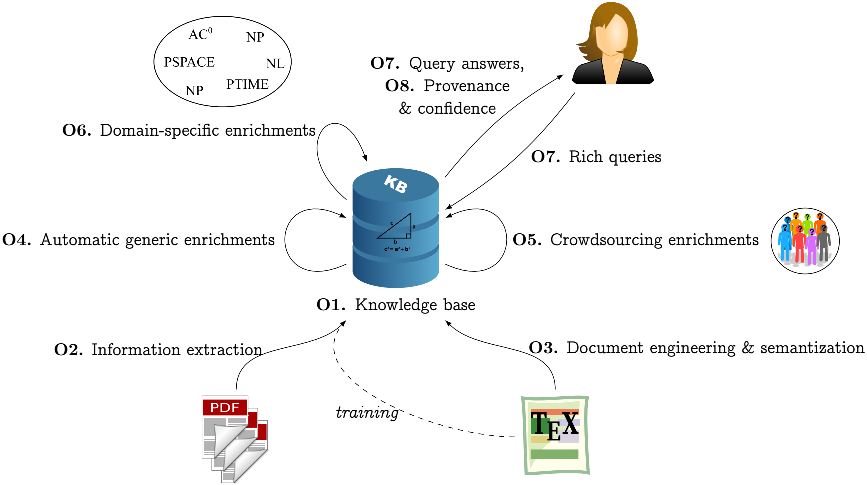
Challenges

- In most cases, only access to PDF versions of scientific articles, with logical structure not reflected in PDFs
- Sometimes (e.g., for the author herself, for arXiv papers), access to (LA)TEX sources, but even those are hard to parse
- Ambiguous references to some theorem of a paper within another
- Ambiguity in the way a theorem is used in the proof of another (relying on it? mentioned as background? actually disproved?)
- Rich background knowledge to be used in specific domains
- Keeping track of the provenance and confidence in the information
- Scalability!

Techniques

- Information extraction from PDF documents
- Semantization and document engineering of (LA)TEX sources
- Training classifiers to disambiguate references
- Incorporating semantic knowledge for specific domains
- Involving human experts and non-experts to help with annotations
- Rich query language and interface
- Open platform, freely available

TheoremKB, A KB of Mathematical Results


AC^0
NP
PSPACE
NL
NP
PTIME
O7. Query answers,
O8. Provenance & confidence
O6. Domain-specific enrichments
O7. Rich queries
KB
$c^2 = a^2 + b^2$
O4. Automatic generic enrichments
O5. Crowdsourcing enrichments
O1. Knowledge base
O2. Information extraction
O3. Document engineering & semantization
PDF
training
TEX

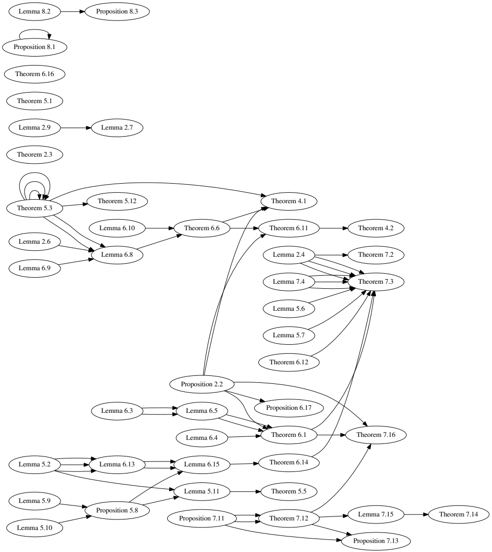
Preliminary results: extraction of theorems

- Detection of boundaries of theorems using:
    - Heuristics
    - Training a Bayes classifier
    - Training a CRF modeling dependencies between a line and the next
- Reasonable performance of CRF's wrt other approaches, even in the absence of geometry information

Research project by Daria Pchelina

Preliminary results: dependency graph

Automatically generated by instrumenting LaTeX commands for theorems and proofs



Merci.

Looking for:

- PhD students
- Post-docs
- Collaborators