JACOBS
UNIVERSITY

Designing Definition Discovery

Read, Recognize, Reflect, Repeat

by

Deyan Ginev

PhD Research Proposal

Prof. Dr. Michael Kohlhase
(Jacobs University)

Prof. Dr. Herbert Jaeger
(Jacobs University)

Dr. Bruce R. Miller
(NIST)

Date of Submission: November 20, 2012

Abstract

Large-scale processing and understanding of scientific, mathematical texts is effectively in "no man's land". Mainstream Computational Linguistics (CL) has so far avoided mathematical discourse for its semi-structured, multisemiotic nature. Knowledge Discovery (KD) methods, while largely applicable, met a lack of available large-scale corpora. Additionally, progress in Mathematical Knowledge Management (MKM) techniques considered informal natural language largely out of scope.

However, we now have big corpora of machine-accessible documents available (arXMLiv, ZBLMath, PlanetMath), as well as a palette of attractive semantic services (e.g. math search, definition lookup), ready to utilize any semanticized corpus. Applying and developing KD techniques on mathematical texts would now enable a dramatic improvement in the quality of e-Learning and the newly emerging Social-Semantic Web both in terms of authoring and document utility. Similar to the explosion of research in the biomedical mining domain, it is time to kick-start a discovery effort for math-rich domains.

This thesis project will make a first step in bridging the fields by looking into the problem of "Definition Discovery" in large-scale collections of scientific documents, namely the automatic classification and annotation of definitional statements and their components. It includes establishing the prerequisites of a systemic foundation and data preprocessing. Taking the discovery process beyond shallow entity mining, we will also attempt a deeper formalization of core statements of interest. There are four independent subproblems:

Read Create a framework for large-scale processing of semi-structured data. Evaluate existing query languages and annotation mechanisms on the two-dimensional sign system of mathematics.

Recognize Tackle the problems of Named Entity Recognition (NER), and Definition Discovery (DD) in scientific domains rich in mathematics.

Reflect Embrace structural semantics to try make explicit the meaning of the near-formal mathematical vernacular employed. Auto-generate a glossary of math entities, capturing the corpus-level structure of mathematical knowledge.

Repeat Experiment with bootstrapping approaches, independently applied to the shallow techniques in the recognition phase and to the deep processing in the reflection phase.

This work is part of a larger vision towards enabling computational understanding or "Machine Reading" of scientific texts. The goal is to approach the AI-hard problem of natural language understanding from datasets already near a formal syntax and semantics, hoping to avoid at least some of the pitfalls of unrestricted human discourse. The added-value of mathematical rigor in scientific writing could turn out to be the key to meeting this challenge.

Contents

1

Introduction and Motivation

The human capacity of language understanding is daunting in its complexity. The multi-dimensionality of the meaning of natural language and the partial and ever elusive regularity of its syntax continue to be an enigma to the scientific community at large. This project proposes a new angle of investigation, by focusing on rigorously written mathematical texts and developing a combination of shallow and deep computational analysis methods. The goal is to enable automatic knowledge acquisition on the object and statement levels, namely discovery of mathematical named entities and definitions, and emerge with a corpus-level ontology of mathematical knowledge.

This chapter offers the preliminary intuitions for the project. Chapter 2 carries out a transdisciplinary survey of the state of art, adapting it into a knowledge acquisition strategy presented in Chapter 3. Chapter 4 concludes by proposing a time plan with a set of milestones, and providing an outlook to the potential impact of the project's success.

1.1 The Nature of "Meaning"

The meaning or "message" of language is multi-dimensional, as it serves an array of functions[1] to fully capture the intention and perspective of the utterer. This is accomplished

[1] some of which **formal**: referential, metalingual and some **experiential**: emotive, poetic, phatic, etc. (see [Jak])

by a complex palette of nuance and vagueness that make the process of understanding more elaborate and reasoning intensive, as well as parametric in the context of the discourse and the background pragmatics of the participants.

It is hence fruitful to focus on scientific articles and especially their mathematical vernacular, which aims at deriving formal truth and hence focuses on the formal functions of language. Indeed, the linguistic distinction between a "word" and a "term" is that terms "*instead of referring to instructions to build points of view, refer to technical concepts*" and are thus "*not supposed to invoke non-deductive inferences*" [Rac00][2]. We proceed onwards with our investigation of meaning, having already seen a hint of the possibly relaxed complexity in the terminology-rich language of scientific texts.

1.1.1 Meaning's Duality – Intrinsic or Extrinsic

It has been long debated in works of Philosophy of Language and Semantics of Natural Language whether and when the denotation of an utterance is self-contained or induced by the context and practice of language use. Respectively, two branches of semantics have sprung up to support both extreme views - classic semantics models a completely intrinsic view on language meaning, while the academically young statistical semantics models an entirely extrinsic one. A common problem is that the experiential dimensions of meaning (such as feelings, ideals, opinions) do not fit in the clear cut frames of logical theories, and vice versa for the cognitive dimensions.

For what it's worth, the author takes a standpoint that recognizes and accepts the duality of language meaning. It seems possible that an intrinsic meaning has an emergent origin as a higher level of complexity over extrinsic meaning, in turn emergent from experience. As the process of perceiving a car approaching fast on a collision course is the result of synthesizing and abstracting over millions of perceived sensory signals (visual, audio, smell, increased heartbeat), conditioned with prior experience, so could be the extrinsic meaning of words. Crucially, one can also produce an intrinsic definition of "a car approaching on a collision course", which will attempt to abstract away over the sensory input that remained secondary and irrelevant to the central notion of "an imminently nearing vehicle posing danger to one's life". This intuition is reaffirmed by the circularity of intrinsic definitions which invariably define signifiers in terms of other

[2] It would be naive to accept that scientific documents do not exhibit any of the other functions of language, as they also partake of social processes, such as submission-reviewing for publishing or teaching-learning for education, in both of which it is essential that the message of the text is presented convincingly in style, form and content. However, they do so in a lesser extent and are ideally sentiment-free in the heart of math vernacular.

signs, always locked inside the universe of human semiosis. The only loophole that allows us to escape this circularity is to assume tacit experience as an extrinsic foundation of meaning.

Similarly, a basic scientific concept such as "number" would emerge from experience as we abstract away from counting real world objects and consequently as we deal with currency in our daily life.[3] However, once the first extrinsic concepts of counting are formed, we would be indoctrinated in their intrinsic counterparts, courtesy of the public education system, and would expand and refine those intrinsic definitions as our study of Mathematics progresses. While the author can immediately relate to an extrinsic nature of counting, it is impossible to claim such a real-world intuition for any advanced numeric construct, e.g. for uncountable infinities, ordinal numbers, irrational and imaginary numbers. For each of these, an act of "imagining" is necessary, which is different from direct experience and much closer to reflection.

It is in this light that I value the rigor of mathematical vernacular and expect to see a clear intrinsic structure to its meaning. Nevertheless, I still accept extrinsic principles for its syntax and vocabulary, as they are naturally emergent from the social exchange and practices of the scientific community. It poses great interest to investigate the interplay between the two paradigms, which becomes most evident exactly in rigorous, yet natural, mathematical vernacular.

Influential works for the author's current position on meaning include Bertrand Russels's "The Analysis of Mind" [Rus21], Sowa's "The Goal of Language Understanding" [Sow12], as well as an illuminating set of discussions during a research visit to TU Braunschweig with prof. Dr. Wolf-Tilo Balke's research group and prof. Dr. Michael Kohlhase.

1.2 The Flexi-Formalist Program

Even when we discard the multi-dimensionality in language function and focus on the cognitive aspects, we still need to face the problem of exposing the knowledge behind the language utterance. As suggested in [KK11], the classic view on the separation between "formal" and "informal" content is too absolute as it fails to facilitate the partial "islands of formality" in human communication. Many functions of communication, and respectively many applications of semantic technologies, require only partial certainty (or "formality") of the communicated message.

[3] Crucially, while we *communicate* of these processes via *natural language*, which acts as the medium of rational thought.

To exemplify, in everyday discourse many dialogues would contain vague informal replies, the semantic payload of which would reside in conveying a "sentiment" or "confirmation/rejection" towards a previously stated proposition, e.g. "It sounds great" or "That might work". In the vernacular of mathematics similar examples can be found with regard to omitting proof steps as "trivial" or "obvious", when the author considers that omission to be irrelevant to the message he's trying to convey. The above examples could demonstrate flexi-formality at the shallow statement-level when given the respective formalizations of "positive reply", "affirmative reply" and "proof step" but remain informal underneath that statement label.

A preliminary classification of what would be "levels of formality" is presented in Figure 1.1. The hourglass metaphor implies a trickling down of semantics, starting from the informal corpus at the top and gradually reaching a fully formalized formal library at the bottom. In Figure 1.1, the basis of the textual scale is the OMDoc ontology of mathematical documents [Koh09]. A generalization of the OMDoc model, the MMT language [RK13] is used for representing the aggregate knowledge bases of deep methods, together with the more classic semantic model of Discourse Representation Theory (DRT) on the semantics side. The choice of using both MMT and DRT was motivated by illustrating the applicability of both models and to emphasize their benefits - while DRT is a very capable model for language semantics, MMT excels at modularity and the management of large-scale formal libraries. Ideally, embedding DRT in the MMT language would provide the best of both worlds. The second distinction to notice is that while the knowledge bases formed by aggregating shallow semantics are distinct artifacts, the bases formed when aggregating real formal semantics remain consistently within the same semantics paradigm, one level up the textual scale. For example, MMT symbols and Discourse Representation Structures (DRS) are largely interchangeable and the distinction between the semantics and the resulting knowledge base becomes blurred.

A flexi-formalist approach to language understanding would gradually build up an ever more formalized version of a target document, starting from its informal human-readable form and continuously enhancing it with partial structural, and hence machine-readable, semantics, opening gateways for new user-assistance applications at every step. It is exciting to investigate the bridge between shallow and deep methods. After the "knowledge quanta" (such as the named entities of mathematical terms) have been detected by shallow KD methods, a pass to the deep methods of CL could continue up on the formalization ladder. Ideally, a symbiosis of shallow and deep methods would provide a semantic "stack trace" leading from the classic "informal" to the classic "formal" end of the flexi-formal semantics spectrum.

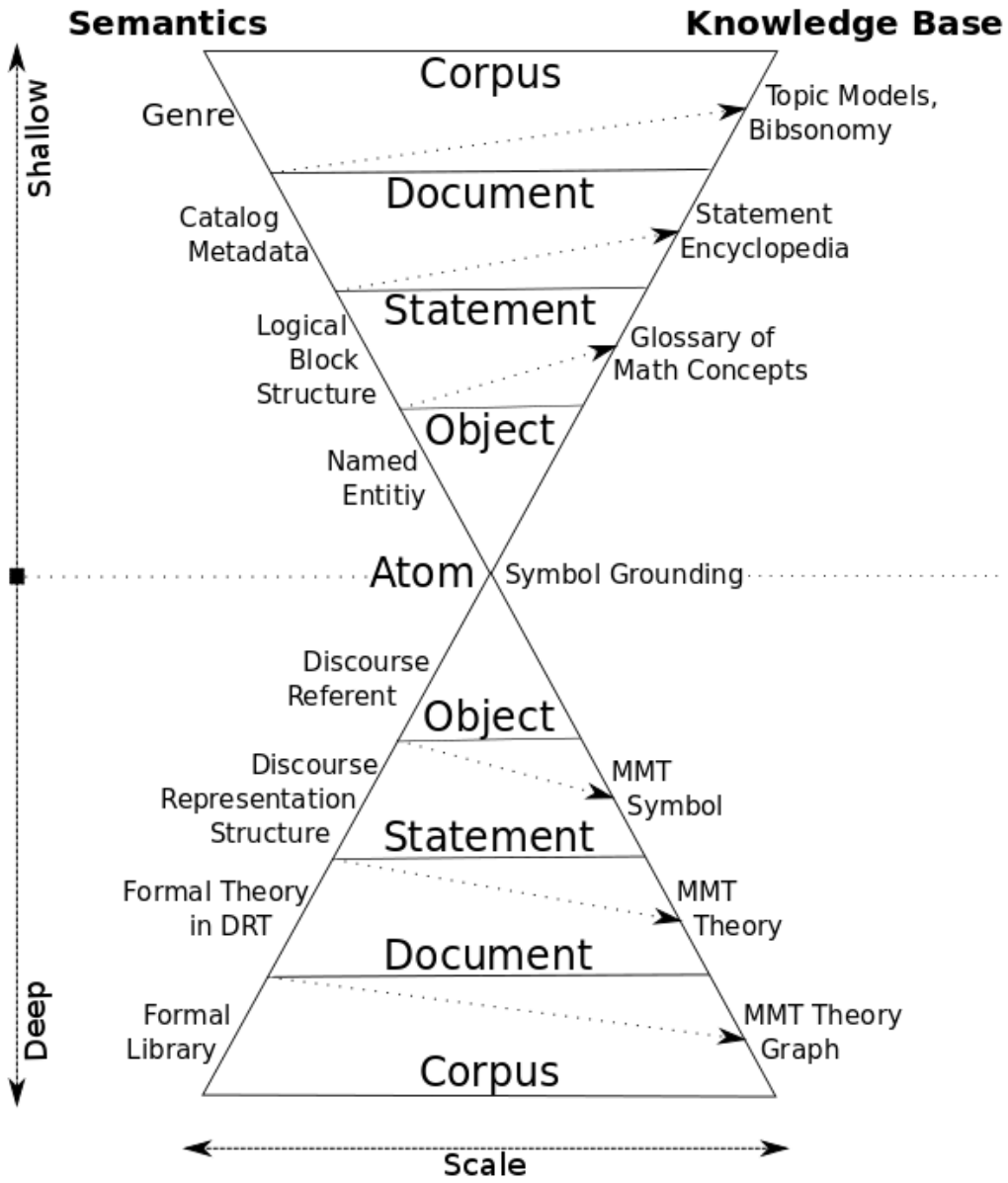


Figure 1.1: Flexi-formality across textual scales

1.3 Knowledge, ubi es?

The term "knowledge" still struggles to find a definition that is universally agreed on. We have named the processes underlying the transmition and formation of knowledge, namely "communication" and "cognition". There is agreement at large on their function, but only partially on their mechanics. On the surface, the process of semiosis enables the transmission of language messages that compositionally encode what we can think of as "knowledge quanta", hidden within. The jury is out on whether semiosis is self-sufficient for human cognition, or additional mnemic structures, or "mental models" facilitate our cognitive processes.

While those concerns fall largely outside of the scope of this project, it is crucial to postulate what "knowledge" is from the perspective of a computational agent, as our purpose lies in the gradual development of machine understanding of language. To this extent, the author takes a flexi-formal view (see Section 1.2) covering the entire semantics spectrum – starting with prerequisite methods rooted in statistical semantics and aiming towards a structural semantics model at the deep end. Tieing the shallow and deep semantics of atoms, i.e. solving the problem of symbol grounding, would hopefully be a successful bridge between the statistical and logical paradigms. In practice, the assumption is that meaning lies in the structure of language utterances, and respectively in structural representations and data types for computational agents, whether statistical or logical.

Taking this mixed approach would induce "knowledge" bases to be pluralistic and heterogeneous. Their definition would be of per-level aggregates of discovered semantics from the original utterances, based on the degree of formality. So we will end up with a range of knowledge bases of various shapes and sizes, each tailored to a different degree of understanding.

1.4 Motivation

It is clear that as humans we will always prefer to use natural language over other, more constrained technologies for communicating with each other. Nevertheless, machine-assistance has proven itself to be a powerful asset in e-Learning (e.g. PlanetMath[Pla], ConneXions [HBK03]), research (Formalized Mathematics, Theorem Proving, Formal Verification), retrieval (Statistical and Semantic Search) and communication (Machine Translation). One array of e-Learning services that is still lagging behind is that of the Active Documents paradigm [HM00], which has the prerequisite of first making a relevant

subset of a document's semantics machine-accessible. Mature Active Document platforms, such as the Planetary system [KCD+11a] would greatly benefit from large-scale deployments, which would also push towards their prime time as Web 3.0 technologies.

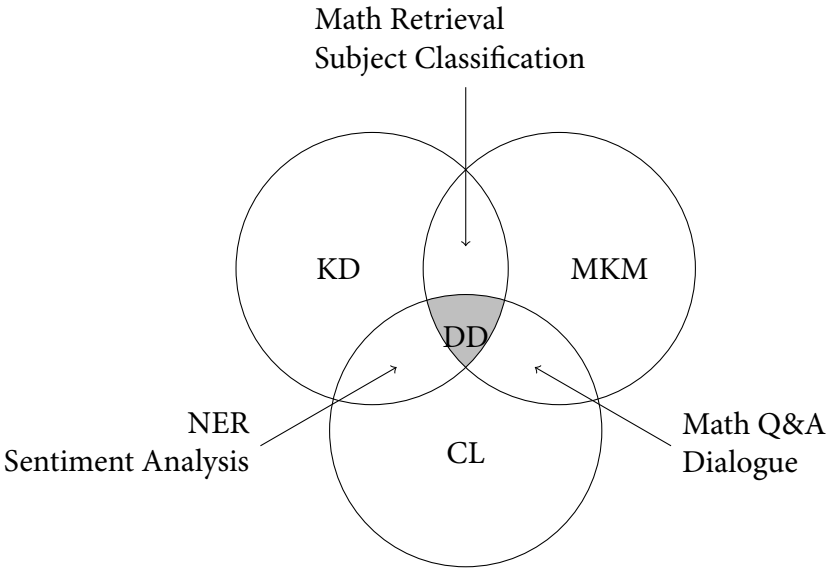


Figure 1.2: Definition Discovery - a transdisciplinary challenge

The Definition Discovery (DD) over mathematical statements, as well as their fine-grained constituents (definiendum and definiens) would immediately feed in an Active Documents framework, e.g. in a semantic service such as definition lookup, giving us a practical incentive to pursue it further. Additionally, the domain of math-rich scientific texts has not been extensively investigated as has been recently described in [Gin11, Chapter 2]. The author recognizes the trans-disciplinary nature of the DD task, as summarized by Figure 1.2. It combines two classic knowledge discovery challenges (classifying definitional statements and named entity recognition), a classic semantics construction task (symbol grounding during knowledge acquisition), a representation task (adapting CL methods with MKM insights for operating on math-rich discourse) and poses foundational questions about the interplay and duality of the semantic models.

The novelty of the core problem at hand is nicely complemented by the availability of the necessary training data[4], as well as bootstrapping methods, making it very promising for producing first results in this direction. Furthermore, any developed approach would

[4]The arXiv corpus has 22,000 explicitly marked up definitions.

be applicable to the general task of large-scale knowledge discovery of statement-level structures, such as axioms, theorems and proofs.

1.4.1 Challenges

While the core task of definition discovery is well-contained and could be seen as a unitary problem in knowledge discovery, in reality it is the tip of an iceberg of prerequisites that need to be met, as shown in Figure 1.3.

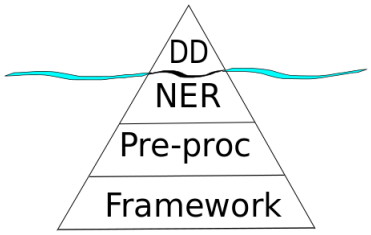


Figure 1.3: Definition Discovery - an Iceberg of Prerequisites

Currently, there are no good technical tools neither for dealing with semi-structured data nor with flexi-formal annotations and representations. This requires adapting one of the existing analysis frameworks for large-scale processing, or alternatively building a new framework from scratch, together with inventing new preprocessing tools and adapting relevant pre-existing analysis techniques to a new domain. Additionally, such framework would require a new capability of automating the representation transitions a flexi-formal process would exhibit, a set of problems that has been outlined in more detail in [GJA+09]. Specifically for Definition Discovery, adapting a solution for Named Entity Recognition would be either a pre- or co-requisite, in order to establish the crucial definiendum and definiens constituents. Any NER method would also need to be extended to cover the semiotic resource of math symbolism, as a lot of defined/defining entities are mathematical expressions, rather than natural language phrases.

2

State of Research

2.1 Language and Mathematics Understanding

The author has previously provided an in-depth overview of the state of Language Understanding [Gin11, Section 2.1], and Mathematical Knowledge Management (MKM) [Gin11, Section 2.2], with respect to work on scientific documents. Taking that as preliminary, the essential summary is that current work on documents rich in mathematics is scarce and young, hence it either lacks in coverage (e.g. [CFK+09]) or does not demonstrate convincing success rates (e.g. [WGK11]).

It is also worth noting that all existing frameworks for large-scale language analysis, such as GATE [CMBT02], UIMA [FL04], or Hadoop [Tay10], are designed for exclusively working with unstructured data, which makes it impossible to directly apply them on the semi-structured data of scientific texts. Even more generally, the flexi-formalist approach to semantization embraces semi-structured data as its lingua franca, as a document artifact spends the majority of its lifespan in intermediate analysis stages, making it paramount to have in-built support for the querying and manipulation of semi-structured data. We investigate these issues in greater detail in Section 3.1.

Currently no scientific community has fully adopted the problem of language and mathematics understanding as their own, and only now the first concerted efforts towards

improving math retrieval and understanding are starting in the NTICR retrieval community [**?**].

2.2 Knowledge Discovery

While knowledge discovery has not specifically targeted math-rich documents, many of the available shallow mining methods of the field are also applicable to their textual fragments. This includes the tasks of Named Entity Recognition (NER) and Definition Discovery (DD). Let us examine a brief sample of the recent best results in the NER task.[1]

[RR09] offers a brief overview of common NER approaches, naming HMM, CRF and sequential applications of Perceptron or Winnow. The paper takes up the CRF approach for its peak performance run, and then looks into four key design decisions in a NER system - the representation of text chunks, the choice of an inference algorithm, the use of external knowledge (gazetteers) and of non-local features. The overall success achieved is a 90.8 F_1 score on the CoNLL03 dataset. Next, [LW09] proposes a different approach, based on clustering of phrases and then using the resulting clusters as features in discriminative classifiers. The phrase clusters are obtained without any labeled data, and are then used to reinforce the training data in a feature-extraction process. Promisingly, the method does not require language-specific linguistic information, such as POS tags, for its optimal performance. That makes it both language-independent and applicable to math documents, where POS-taggers trained over newswire texts perform poorly. This approach reaches 90.9 F_1 score on the CoNLL03 dataset.

[PLB$^+$06] focuses on a bootstrapping-based approach for fact extraction, suitable for scenarios where no available training data is present. In Definition Discovery, the definiendum and definiens constituents could be seen both as named entities to be recovered by NER, or as introduced facts that could be extracted on the basis of binary relations between the two named entities. That makes the methods of [PLB$^+$06] applicable, especially given the lack of any training data for definienda and definiens in mathematical corpora. With as few as 10 seed facts, the paper accomplishes a 90% precision extraction of one million facts of a given type.

Taking bootstrapping approaches a step further, CMU's Never-ending Language Learning (NELL) [CBK$^+$10] project attempts an endless reiteration of knowledge acquisition and learning, while building a consistent knowledge base of all learnt facts. NELL is very

[1] The author thanks Cevahir Demirkiran from ZBLMath, for his pointers to relevant literature.

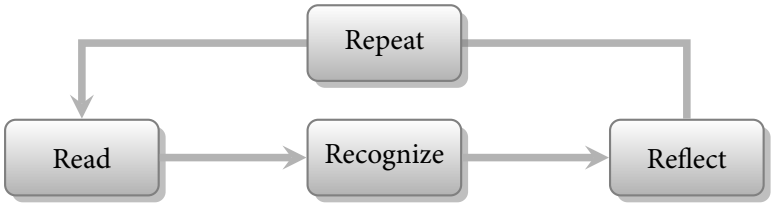
close to the overall goal of this thesis project and the author is excited to try out their approach in a flexi-formal setting of reusable framework components. As a note of caution, one of the boons of bootstrapping-based approaches is the impossibility of realistically evaluating recall, as it is common to work on enormous datasets deprived of any training data.

In the meantime, the MKM community has been developing tools independently, a noteworthy one of which is the NNexus [GKX09] auto-linker. NNexus links any given data entry (in HTML) against a manually curated gazetteer of known concepts. It performs no discovery analysis in its own right, as it uses a simple "longest phrase match" heuristic for classification and relies on the system's users to manually discard any false positives. This serves as an important contrast between the KD and MKM communities, the former of which have focused on developing ever better discovery methods, while the latter have invested in creating user-friendly systems that manage the semi-structured representations of scientific documents.

In conclusion, there is fertile ground for adapting and combining methods from the KD and MKM communities, as well as hope for utilizing classic CL methods for the limited scope of mathematical vernacular.

3

Approach and Methods



The proposed approach targets the creation of a complete knowledge acquisition pipeline, centered around the task of Definition Discovery. As such, the main goal is to face the challenge of DD in the real-world setting of Big data and with the vision of bringing the analysis effort to the international community, by allowing standardized distributed workflows.

3.1 Read

The author has available the data of the PlanetMath [Pla](≈ 8000 encyclopedia entries), arXiv [ArX]($\approx 800,000$ scientific articles) and ZBLMath [ZBM]($\approx 3,000,000$ math reviews) corpora, as well as a custom collection of 15 freely licensed textbooks in mathematics. In the arXMLiv project [SKG+10], we have seen that it takes between one and two processor years to do the basic representation preprocessing over the arXiv corpus

alone, making the need for a distributed processing setup clear from the very start. As actual analysis on top of these corpora would be significantly more intensive, the need for tractable distributed workflows becomes even more apparent.

Hence, the author is faced with the need to adapt or create a distributed framework which can handle process jobs on several million entries with an overall estimated size of 1-5 TB, including the expected annotations and newly generated resources. An additional constraint is the reality of available hardware, namely 40 basic desktop units.[1]

I will consider two possible lines of attack. One is the classic monolithic design, with a central conductor controlling a number of distributed or federated backends. Think of a web application behind a load-balancer remotely communicating to its backend clusters, as a central point of contact. The backends in mind are an XML database, capable of efficiently manipulating semi-structured data, as well as a Semantic Web triple store, ideal for interoperable stand-off annotations typical of shallow analysis methods. The advantages of this approach are rooted in the loose coupling of components, allowing for easy adaptation of pre-existing tools and for minimal restrictions for distributed external components. In particular, external contributors would be able to use their programming language of choice and volunteer their own processing hardware. In this scenario, the central-access backends become the system bottleneck.

The alternative approach, clearly superior when facing truly Big data (petabytes and up), would be to use a general-purpose distributed framework such as Hadoop, deployed on some computational cluster. The restriction comes in the hard-coupling of components - all processing must run on the same cluster, as well as all contributions need to be written in the same language. An additional restriction is the simplicity of the MapReduce paradigm, which currently has no production-ready advanced query languages, such as XQuery or SPARQL, which would be available in a monolithic approach.

Given these constraints, a compromise might be necessary to ensure either the short-term success of the discovery methods in this project[2] or the long-term utility of the created framework[3]. In either case, the individual steps to realizing definition discovery (representation purification, tokenization, statistic models) would be realized as separate modules conforming to a standard framework API. The mentioned API should be minimal, yet useful, and the framework should provide a basic dependency management that would allow for the community development of analysis tools.

[1] A cloud-based solution is not yet ruled out, however, which might prove a viable outsourcing of concerns.

[2] It is quite possible that all necessary methods can be realized in the MapReduce framework

[3] Any slightly complex analysis technique would quickly come to require advanced query languages, in order to fully utilize semi-structured data.

An orthogonal necessity would be a representation-awareness of the API, as different methods would require different data structures. Whether built-in or provided as separate modules, methods for representation transitions will also be provided as part of this project, also following the transition model in [GJA+09].

The success of the design and implementation of these ideas will be measured by the brevity and ease of development of the various analysis components.

3.2 Recognize

Definition 1. *In the following, let us call **declarations** phrase or sentence-level language constructs that declare one term, a **definiendum**, in terms of a language expression, a **definiens**. In KD, a declaration is assumed to introduce a "fact" that could later be the target of a "fact extraction" discovery challenge.*

Definition 2. *We call a **definition**, a statement-level construct, e.g. a paragraph, of mathematical importance, containing one or more declarations. A definition is thus a deeper property than a declaration as it asserts some prominence of the declared object(s).*

The previous two paragraphs contain examples of declarations[4] and are in themselves examples of definitions. The core task of this thesis is to recognize definitions and their constituent declarations in order to provide added-value machine-accessible semantics.

3.2.1 NER

Thus, one has to solve the NER problem in mathematical texts. An interesting research question is how to adapt the existing NER methods to the additional semiotic resource of mathematical expressions. One approach is to normalize the mathematics away to named entity artefacts, serving as unique identifiers, and then try the classic solutions as-is.

Two issues that need to be addressed, however, are the lack of training data for declarations and the loss of information from the constituent structure of math expressions, which could play a crucial role in determining the term-hood of a formula. For example, relational operators would not be common in term names, except in scripts of a special

[4] The definienda are marked in bold.

form and only decorating alphabetic symbols, as opposed to big operators, such as integral and sum. In short, there is a rich feature space to be explored inside the structure of expressions. As all high-performing NER results rely on some form of gazetteer, it will also be interesting to investigate math-enhanced gazetteers.

3.2.2 Definition Discovery

To contrast with declarations, the arXiv corpus has roughly 22,000 explicitly marked up definition environments, that can be used as training data for semi-supervised classification approaches. We could think of this as a binary classification task for logical paragraphs, where each of the marked up paragraphs falls in the class of definitions, while each of the regular, non-definitional paragraphs (crucially from the same documents as the marked up ones) falls in the class of non-definitions.

It is again the case that it would be fruitful to adapt existing approaches, such as HMM and CRF, paying close attention to the preprocessing of math expressions. It is also interesting to consider the interplay between DD and NER tasks, as the results of each one could be leveraged as features for the other.

3.3 Reflect

Recall the flexi-formal hourglass in Figure 1.1. We have looked in one statement-level task (DD) and one object-level task (NER), proceeding downwards to the atomic middle-ground. The scope of this thesis extends only to the examination of the basic bridge between the shallow and deep paradigms, in the challenge of symbol grounding.

Ideally, the NER and DD tasks would provide us with contextualized declarations, within some definition of some certain document. One would next want to record this information to create a single knowledge graph that can be later used for added-value services. Such a step would typically be called creating a glossary of known terms.

However, while the defined terms might be unique and clearly related on a document-level, that is far from being the case on the corpus-level. The **symbol grounding** problem poses a challenge to identify when two symbols which carry the same name carry a different meaning, and vice versa - when to symbols that carry different names share a common meaning (or in our rigorous case - an equivalent definiens).

Both statistic and classic semantics have answers to what symbol equivalence should be, and given that we are investigating real-world math definitions, the author would be leaning towards taking up the formal approach in this case. In other words, the goal would be to fully formalize the definiens of the DD- and NER-recognized symbols and create a corpus-level ontology of their interrelations. As each symbol name is uniquely identified by its textual location, we have a good fit with the MMT paradigm of modular theories, each document standing in for a theory. In that light we can say that the recognition methods discussed in the previous section would establish the declarations, and hence the entirety of would be MMT theories, while the reflection step would derive partial views between those theories, aiming to identify equivalent symbols and respectively subexpressions.

We thus climb the semantic ladder from informal strings, through gazetteers, glossaries, ontologies and reaching the modular landscape of MMT theories. As a proof-of-concept pipeline through most of the flexi-formal spectrum is a central goal of the proposed research, accomplishing this step on a non-trivial dataset would be considered a big step forward.

3.4 Repeat

There are big similarities between working with community corpora such as arXiv and with web-as-corpus data. Both were never custom curated by linguists, both contain noise and spurious errors, both lack any significant training datasets or gold standards. Hence, the methods that perform best with web-as-corpus tasks, namely bootstrapping techniques, seem very likely candidates for our collections of scientific documents. Such methods could be either counterparts or substitutes to the recognition methods in Section 3.2, especially since once annotated, the discovered data by a bootstrapping approach could be in turn used as a training set for a different recognition algorithm.

On the shallow end, given representation preprocessing, pre-existing methods could be used as-is, given a small starting collection of seed expressions for declarations[5]. The techniques based on binary relation patterns of the form "*pre* ***hyponym*** *relation* ***hypernym*** *post*", could be identically applied on declarations patterns, i.e. "*pre* ***definiendum*** *relation* ***definiens*** *post*".

[5]For example: "*let* ***definiendum*** *be a* ***definiens***" or "*we call* ***definiendum definiens***"

On the deep end, the NELL project [CBK+10] is investigating the automatic construction of ontologies, iterative refinement of concepts[6]. It is unclear to what extent the NELL methods are applicable when math expressions become involved, as it is possible to confuse notation variants with the need to refine a concept, or similarly for definitional variants (e.g. the natural numbers defined to include or exclude 0).

An interesting difference between web-as-corpus data and scientific articles is that the analysis process could leverage a well-defined perspective in reading. While data on the web has all of its prerequisites implicit, scientific articles have explicit bibliographies that define a type of dependency graph between articles.

[6] See Read the Web (`http://rtw.ml.cmu.edu/rtw/`) for current status.

4

Goals and Milestones

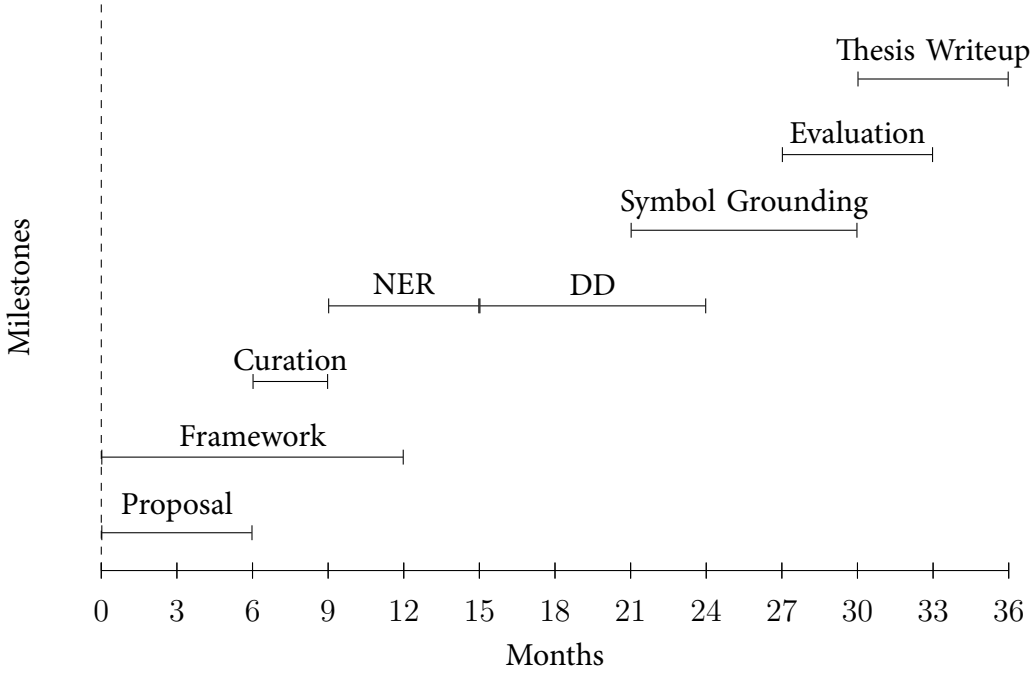
4.1 Goals

My main goal in this research effort is to break ground in the transdisciplinary challenge of Definition Discovery (DD), using a flexi-formalist approach. A secondary goal is to successfully create a bridge for the scientific communities in Computational Linguistics, Knowledge Discovery and Mathematical Knowledge Management that would enable fruitful future collaborations between the fields. To accomplish this, a three-fold success would be necessary.

First, a production-ready *large-scale analysis framework* must be available both as deployment and toolkit for the scientific community at large. It will offer a solution to distributed workflows and collaborations, as well as to the scalability challenges of Big data. Second, a full *analysis pipeline* built from the ground-up performing Definition Discovery must be developed and evaluated. Finally, the *core problem of DD* should be theoretically grasped, and a better understanding of the techniques and trade-offs in bridging statistical and formal methods must be presented.

The main applications benefiting from such *semanticizing of large-scale corpora* of scientific texts, particularly their mathematical vernacular, would be services in the Active Document paradigm (e.g. definition lookup in [KCD+11b]), as well as semantic approaches to Math Information Retrieval (e.g. "applicable theorem search" in [KŞ07]).

4.2 Timeplan and Milestones



1. *Proposal* - specify a project attack plan by completing the current document
2. *Framework* - survey the state of art and create or adapt a framework for large-scale processing on semi-structured datasets.
3. *Curation* - prepare training datasets, transform data to appropriate representations, tackle common preprocessing tasks
4. *Named Entity Recognition (NER)* - Evaluate and adapt existing NER methods, including bootstrapping-based approaches.
5. *Definition Discovery (DD)* - Evaluate and adapt classification methods for definitional paragraphs, including bootstrapping-based approaches.
6. *Symbol Grounding* - Attempt formalization of sentence-level declarations, grounding and interlinking their definienda in the global corpus context.
7. *Evaluation* - Final evaluation of the chosen methods.
8. *Thesis Writeup* - Finalize project work by delivering a final thesis document.

Bibliography

[ArX] arxiv.org e-Print archive. web page at http://www.arxiv.org. seen November 2012.

[CBK+10] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.

[CFK+09] Marcos Cramer, Bernhard Fisseni, Peter Koepke, Daniel Kühlwein, Bernhard Schröder, and Jip Veldman. The naproche project controlled natural language proof checking of mathematical texts. In Norbert E. Fuchs, editor, *CNL*, number 5972 in Lecture Notes in Computer Science, pages 170–186. Springer, 2009.

[CMBT02] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the ACL*, 2002.

[FL04] David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348, 2004.

[Gin11] Deyan Ginev. The structure of mathematical expressions. Master's thesis, Jacobs University Bremen, Bremen, Germany, August 2011.

[GJA+09] Deyan Ginev, Constantin Jucovschi, Stefan Anca, Mihai Grigore, Catalin David, and Michael Kohlhase. An architecture for linguistic and semantic

analysis on the arXMLiv corpus. In *Applications of Semantic Technologies (AST) Workshop at Informatik 2009*, 2009.

[GKX09] J Gardner, A Krowne, and L Xiong. NNexus: An Automatic Linker for Collaborative Web-Based Corpora. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):829–839, 2009.

[HBK03] Geneva Henry, Richard G. Baraniuk, and Christopher Kelty. The ConneXions project: Promoting open sharing of knowledge for education. In *Syllabus, Technology for Higher Education*. 2003.

[HM00] E. Heinrich and H. Maurer. Active documents: Concept, implementation, and applications. *Journal of Universal Computer Science*, 6(12):1197–1202, 2000.

[Jak] Jakobson's functions of language. online at `http://en.wikipedia.org/w/index.php?title=Jakobson%27s_functions_of_language&oldid=520624390`. seen November 2012.

[KCD+11a] Michael Kohlhase, Joe Corneli, Catalin David, Deyan Ginev, Constantin Jucovschi, Andrea Kohlhase, Christoph Lange, Bogdan Matican, Stefan Mirea, and Vyacheslav Zholudev. The planetary system: Web 3.0 & active documents for stem. In *Procedia Computer Science* [KCD+11b], pages 598–607. Finalist at the Executable Paper Grand Challenge.

[KCD+11b] Michael Kohlhase, Joe Corneli, Catalin David, Deyan Ginev, Constantin Jucovschi, Andrea Kohlhase, Christoph Lange, Bogdan Matican, Stefan Mirea, and Vyacheslav Zholudev. The planetary system: Web 3.0 & active documents for stem. *Procedia Computer Science*, 4:598–607, 2011. Finalist at the Executable Paper Grand Challenge.

[KK11] Andrea Kohlhase and Michael Kohlhase. Towards a flexible notion of document context. In *Proceedings of the 29th annual ACM international conference on Design of communication (SIGDOC)*, pages 181–188, New York, NY, USA, 2011. ACM Special Interest Group for Design of Communication, ACM Press.

[Koh09] Michael Kohlhase. OMDoc: An open markup format for mathematical documents; language specification, primer, projects, applications [version

1.6 (pre-2.0)]. Draft Specification `https://svn.omdoc.org/repos/omdoc/trunk/doc/spec/main.pdf`, 2009.

[KŞ07] Michael Kohlhase and Ioan Şucan. System description: MathWebSearch 0.3, a semantic search engine. 2007.

[LW09] D. Lin and X. Wu. Phrase clustering for discriminative learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1030–1038. Association for Computational Linguistics, 2009.

[Pla] PlanetMath.org – math for the people, by the people. `http://www.planetmath.org`. seen March 2013.

[PLB+06] M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1400. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[Rac00] Pierre-Yves Raccah. Lexical and dynamical topoi in semantic description: A theoretical and practical differentiation between words and terms. In *Language, Text, and Knowledge: Mental Models of Expert Communication*, pages 11–30, 2000.

[RK13] Florian Rabe and Michael Kohlhase. A scalable module system. *Information & Computation*, pages 1–95, 2013.

[RR09] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[Rus21] Bertrand Russel. *The Analysis of Mind*. Library of philosophy. G. Allen & Unwin, 1921.

[SKG+10] Heinrich Stamerjohanns, Michael Kohlhase, Deyan Ginev, Catalin David, and Bruce Miller. Transforming large collections of scientific publications to XML. *Mathematics in Computer Science*, 3(3):299–307, 2010.

[Sow12] John F. Sowa. The Goal of Language Understanding. online at `http://www.jfsowa.com/talks/goal.pdf`, 2012. seen 2012/11/12.

[Tay10] Ronald Taylor. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics*, 11(Suppl 12):S1+, 2010.

[WGK11] Magdalena Wolska, Mihai Grigore, and Michael Kohlhase. Using discourse context to interpret object-denoting mathematical expressions. In Petr Sojka, editor, *Towards Digital Mathematics Library, DML workshop*, pages 85–101. Masaryk University, Brno, 2011.

[ZBM] Zentralblatt MATH. web page at `http://www.zentralblatt-math.org/zbmath`. seen November 2012.