CEUR-WS.org/Vol-2523/paper28.pdf

Towards a Semantically Annotated Corpus of Educational Mathematical Texts in Russian

Olga Nevzorova[1,2], Alexander Kirillovich[1],
Konstantin Nikolaev[1], and Kamilla Galiaskarova[1]

[1] Kazan Federal University, Kazan, Russia
[2] Tatarstan Academy of Sciences, Kazan, Russia

onevzoro@gmail.com, alik.kirillovich@gmail.com,
konnikolaeff@yandex.ru, galias-alsu@yandex.ru

Abstract. We discuss a semantically annotated corpus of educational mathematical texts in Russian. The objective of our research is to create a test collections for automatic formalization of educational mathematical documents. The corpus includes mathematical assertions extracted from educational math textbooks. We manually annotated each assertion as the formula representation in LaTeX and created the formalization of the formula in OpenMath. Symbols used in OpenMath representations are defined in OntoMath^{Edu}, a new educational mathematical ontology.

Keywords: Mathematics, Corpus, Ontology, OpenMath, OntoMath^{Edu}.

1 Introduction

Most of mathematical knowledge is currently recorded in the form of informal documents, consisting of natural language text mixed with formulas in presentation markup. The meaning of such documents is accessible to human readers, but not to machines. In order to this meaning can be machine-actionable, the documents have to be formalized and represented in a form that computers can act on. In practice, full formalization is not necessary, and in fact representation of same semantics only can be enough. This "flexiformalization" paves the way to intelligent mathematical knowledge management applications such as semantic search services, recommender systems, etc. [1, 2]

We study the math assertions in math textbooks for secondary schools. Many of such assertions have the form of plain natural language text but not math statements on formal math language. Our objective is to create a translator of math assertions represented in the form of natural language text to formula representations. These representations we are planning to use in content markup. This development, in turn, requires training and test collections.

In this paper we consider an experimental semantically annotated math corpus, that consists of math assertions extracted from educational math documents. Each asser-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

tion is manually annotated as the formula representation in LaTeX and later we create the formalization of this formula in OpenMath [3]. Symbols used in OpenMath representations are defined in OntoMath^{Edu} (https://github.com/CLLKazan/OntoMathEdu), a new educational mathematical ontology [4]. We believe that this ontology will serve as a Linked Open Data hub for mathematical education. Concepts of the ontology contain labels in English, Russian and Tatar and will be interlinked with the external lexical resources from the Linguistic Linked Open Data (LLOD) cloud [5], first of all, WordNet [6], BabelNet [7], RuThes Cloud [8] and Russian-Tatar Thesaurus [9].

The rest of the paper is organized as follows. In Section 2, we briefly review some projects of building formal and informal mathematical corpora. In Sections 3 we describe the corpus and the process of its construction. In conclusion, we outline the directions of future work.

2 Related Works

In this section we briefly describe informal, formal and parallel informal/formal mathematical corpora.

Informal corpora. arXiv (https://arxiv.org/) is the largest informal mathematical corpus in the world. Its content is represented in LaTeX format. arXMLiv (https://kwarc.info/projects/arXMLiv/) [10] contains arXiv collection, automatically converted to XML, HTML 5 and Content MathML, and making it is more suitable for machine processing.

Formal corpora. The Mizar Mathematical Library (http://mizar.uwb.edu.pl/library/) is the largest corpus of fully formalized mathematics.

Parallel informal/formal corpora. One of the largest manually-created parallel informal/formal corpora is based on the Flyspeck Project. Flyspeck [11] (https://github.com/flyspeck/flyspeck) is a project, which gives a formal proof of the Kepler conjecture in the HOL Light proof assistant. This project is based on the informal book [12] in LaTeX. Approximately 500 formal statements have been aligned with their informal counterparts. The corpus is available by a user-friendly wiki interface [13].

In [14] Kaliszyk et al. lunched a project aimed at automatic translation of informal mathematical texts into formal ones on base of machine learning methods, trained on aligned informal/formal mathematical corpora. In the subsequent works they pesented several synthetic informal/formal corpora as well as translators trained on them. For example, in [15] they presented a neural network translator from informalized LaTeX-written Mizar texts into the formal Mizar language. The training corpus has been generated by transformation of Mizar to natural language LaTeX text on the basis of the existed method developed for presenting the Mizar articles in the journal *Formalized Mathematics*. In [16, 17] they presented a system for parsing ambiguous formulas from the Flyspeck project. The training informal/formal corpus has been constructed by ambiguation of formal statements from the HOL Light theorems in Flyspeck.

The Formal Abstracts (https://formalabstracts.github.io/) is ongoing project, aiming at formalization of the main results of informal mathematical documents (for example, formalization of the main theorem of a research paper). This formalization is also intended to be used in machine learning tasks.

For our knowledge, there is not neither parallel informal/formal mathematical corpus for Russian nor parallel educational mathematical corpus, so the development of such corpus is needed.

3 Corpus description and construction

The corpus is organized as a collection of records. Each record includes the following three fields:

- Russian sentence, extracted from educational textbooks.
- Formula representation of this statement in LaTeX format.
- Formalization of this formula in OpenMath format, where OntoMath^{Edu} ontology is used as an OpenMath content dictionary.

When building the corpus, the following tasks are successively solved.

3.1 Natural language statements extraction

At the first step, we manually extract Russian sentences from education textbooks. We use the secondary school geometry books for 7th–9th grades. The extracted statements are classified according to the following simple classification scheme:

- Class 1: Statements of equality

  a. with complex statement in the left part and simple right part (e.g. positive integer). Example: "The sum of the degree measures of two acute angles of a right triangle is 90°".
  b. with comparison between equivalent components. Example: "The area of a rectangle is equal to the product of its adjacent sides".

- Class 2: Statements of inequality. Example: "Each side of a triangle is less than the sum of two other sides".
- Class 3: Definitions of mutual arrangement (e.g. perpendicularity). Example: "The diagonals of a square are mutually perpendicular".
- Class 4: Composite statements (several formulas in one statement provided with "AND" preposition). Example: "The middle line of a trapezoid is parallel to the its bases and equal to their half-sum".
- Class 5: Conditional statements. Example: "If the angle of one triangle is equal to the angle of another triangle, then the ratio of the area of one triangles to the area of another triangle is equal to the ratio of the product of the sides, enclosing equal angles of one triangle to the product of such sides of another triangle".

3.2 Statements explication

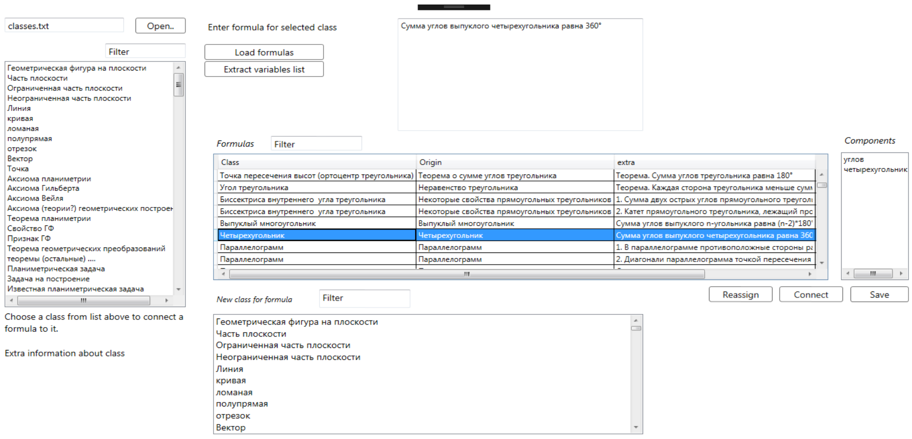
In the extracted statements, many concepts are mentioned only implicitly due to metonymy, ellipsis, etc. For example, for the statement “The sum of the angles of a convex n-gon is (n-2) × 180°” it is assumed that the units of measurement for angles are used in this sum, rather than the angles themselves. Therefore, in the second stage we explain implicit concepts in the extracted statements. Table 1 contains examples of original statements and their explanations.

Table 1. Examples of statements explication

Original (Russian)	Explicated (Russian)	Original (English)	Explicated (English)
Сумма углов выпуклого n-угольника равна (n-2) × 180°	Сумма градусных мер углов выпуклого n-угольника равна (n-2) × 180°	The sum of the angles of a convex n-gon is (n-2) × 180°	The sum of the degree measures of the angles of the convex n-gon is (n-2) × 180°
Средняя линия трапеции параллельна основаниям и равна их полусумме	Средняя линия трапеции параллельна её основаниям и её длина равна полусумме длин оснований	The middle line of the trapezoid is parallel to the bases and equal to their half-sum	The middle line of the trapezoid is parallel to its bases and its length is equal to half the sum of the base lengths

3.3 Concepts annotation

At the third step, we annotate math concepts in the extracted statements. The concepts are annotated in terms of OntoMath^{Edu} ontology. For example, the statement “The middle line of the trapezoid is parallel to the bases and equal to their half-sum” contains the following classes of OntoMath^{Edu} ontology: *Middle line*, *Trapezoid*, *Base*, etc. The tool for this annotation is represented at Fig. 1.



classes.txt

Enter formula for selected class

Сумма углов выпуклого четырехугольника равна 360°

Геометрическая фигура на плоскости
Часть плоскости
Ограниченная часть плоскости
Неограниченная часть плоскости
Линия
кривая
ломаная
полупрямая
отрезок
Вектор
Точка
Аксиома планиметрии
Аксиома Гильберта
Аксиома Вейля
Аксиома (теории?) геометрических построен
Теорема планиметрии
Свойство ГФ
Признак ГФ
Теорема геометрических преобразований
теоремы (остальные) ....
Планиметрическая задача
Задача на построение
Известная планиметрическая задача

Choose a class from list above to connect a formula to it.

Extra information about class

Class	Origin	extra
Точка пересечения высот (ортоцентр треугольника)	Теорема о сумме углов треугольника	Теорема. Сумма углов треугольника равна 180°
Угол треугольника	Неравенство треугольника	Теорема. Каждая сторона треугольника меньше сумм
Биссектриса внутреннего угла треугольника	Некоторые свойства прямоугольных треугольников	1. Сумма двух острых углов прямоугольного треугол
Биссектриса внутреннего угла треугольника	Некоторые свойства прямоугольных треугольников	2. Катет прямоугольного треугольника, лежащий прс
Выпуклый многоугольник	Выпуклый многоугольник	Сумма углов выпуклого n-угольника равна (n-2)*180°
Четырехугольник	Четырехугольник	Сумма углов выпуклого четырехугольника равна 360
Параллелограмм	Параллелограмм	1. В параллелограмме противоположные стороны ра
Параллелограмм	Параллелограмм	2. Диагонали параллелограмма точкой пересечения

углов
четырехугольник

Геометрическая фигура на плоскости
Часть плоскости
Ограниченная часть плоскости
Неограниченная часть плоскости
Линия
кривая
ломаная
полупрямая
отрезок
Вектор

Fig. 1. GUI of the concept annotation tool

3.4 Representation of statements as formulas

At the next stage, we represent the statements as the formulas in LaTeX. Table 2 contains examples of this representation as formula statements.

Table 2. Examples of statements and its formula representation

Statement (Russian)	Formula representation (Russian)	Statement (English)	Formula representation (English)
Сумма углов выпуклого n-угольника равна (n-2) ×180°	$\angle A_1 + \angle A_2 + \ldots + \angle A_n = (n\text{-}2) \times 180°$, где $A_1 \ldots A_n$ – выпуклый n-угольник; $\angle A_1, \angle A_2, \ldots, \angle A_n$ – углы выпуклого n-угольника	The sum of the angles of a convex n-gon is (n-2) × 180°	$\angle A_1 + \angle A_2 + \ldots + \angle A_n = (n\text{-}2) \times 180°$, where $A_1 \ldots A_n$ is a convex n-gon; $\angle A_1, \angle A_2, \ldots, \angle A_n$ are the angles of this convex n-gon
Сумма двух острых углов прямоугольного треугольника равна 90°	$\angle ABC + \angle BAC = 90°$, где ABC – прямоугольный треугольник; $\angle BCA$ – прямой угол	The sum of the two acute angles of a right triangle is 90°	$\angle ABC + \angle BAC = 90°$, where ABC is a right triangle; $\angle BCA$ is a straight angle

3.5 Formalization of the formulas in OpenMath

At the final step, we formalize formulas in OpenMath format. We use OntoMath^{Edu} ontology as a content dictionary in this formalization.

4 Conclusion

In this paper, we presented a semantically annotated corpus of educational math texts in Russian. The corpus consists of natural language statements, extracted from an educational textbook. Extracted statements were manually complemented by its representation as LaTeX formulas and OpenMath formal representation. As a OpenMath content dictionary we used OntoMath^{Edu} ontology.

The corpus now is still on the development stage, so our immediate goal is to release the first working version.

After that we are going to adopt it in the development of the components of a new digital educational platform, which is intended for solving such tasks as automatic knowledge testing; automatic recommendation of educational materials according to an individual study plan; and semantic annotation of educational materials. In particular, the corpus is intended to be used for training an automatic translator from Russian educational documents to its formal representation, as well as a test collection for an ontology-based mathematical information extraction tool. Also, the corpus can be used to verbalize a formal mathematical document as a natural language text in Russian. Additionally, we are going to use it for enrichment of OntoMath^{Edu} ontology.

The corpus will be published at the Linked Open Data (LOD) cloud.

Acknowledgements

This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement no. 1.2368.2017, and Russian Foundation for Basic Research and the government of the region of the Russian Federation, grant № 18-47-160007.

References

1. Kohlhase, M.: The Flexiformalist Manifesto. In: Voronkov, A., et al. (eds.) Proceedings of the 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2012), pp. 30-35. IEEE (2012). doi:10.1109/SYNASC.2012.78
2. Kohlhase, A. and Kohlhase, M.: Towards a Flexible Notion of Document Context. In: Protopsaltis, A., et al. (eds.) Proceedings of the 29th ACM international conference on Design of communication (SIGDOC 2011), pp. 181-188. ACM (2011). doi:10.1145/2038476.2038512
3. Buswell, S., Caprotti, O., Carlisle, D. P., Dewar, M. C., Gaëtano, M., and Kohlhase, M.: The OpenMath Standard, Version 2.0. The OpenMath Society (2004). https://www.openmath.org/standard/om20-2004-06-30/omstd20.html
4. Kirillovich, A., Nevzorova, O., Falileeva, M., Lipachev, E., Shakirova, L.: OntoMath^{Edu}: Towards an Educational Mathematical Ontology. In: Kaliszyk, C., et al. (eds.) Workshop Papers at 12th Conference on Intelligent Computer Mathematics (CICM-WS 2019). CEUR Workshop Proceedings (forthcoming)
5. McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., and Pool, J.: The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In: Calzolari N., et al. (eds.) Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 2435-2441. ELRA (2016).
6. McCrae, J. P., Fellbaum, C., and Cimiano, P.: Publishing and Linking WordNet using lemon and RDF. In: Chiarcos C. et al. (eds.) Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014), pp. 13–16. ELRA (2014)
7. Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., and Navigli, R.: Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In: Calzolari N., et al. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 401–408. ELRA (2014)
8. Kirillovich, A., Nevzorova, O., Gimadiev. E., and Loukachevitch, N.: RuThes Cloud: Towards a Multilevel Linguistic Linked Open Data Resource for Russian. In: Różewski, P. and Lange, C. (eds.) Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017). Communications in Computer and Information Science, vol. 786, pp. 38-52. Springer, Cham (2017). doi:10.1007/978-3-319-69548-8_4
9. Galieva, A., Kirillovich, A., Khakimov, B., Loukachevitch, N., Nevzorova, O., and Suleymanov, D.: Toward Domain-Specific Russian-Tatar Thesaurus Construction. In: Proceedings of the International Conference IMS-2017, pp. 120–124. ACM (2017). doi:10.1145/3143699.3143716
10. Stamerjohanns, H., Kohlhase, M., Ginev, D., David, C., and Miller, B.: Transforming Large Collections of Scientific Publications to XML. Mathematics in Computer Science, 3(3), 299–307 (2010). doi:10.1007/s11786-010-0024-7

11. Hales, T. C.: Introduction to the Flyspeck project. In: Coquand, T., Lombardi, H. and Roy, M.-F. (eds.) Mathematics, Algorithms, Proofs. Dagstuhl Seminar Proceedings, vol. 05021. IBFI (2006)
12. Hales, T.: Dense Sphere Packings: A Blueprint for Formal Proofs. Cambridge University Press (2012)
13. Tankink, C., Kaliszyk, C., Urban, J., and Geuvers, H.: Formal mathematics on display: a wiki for Flyspeck. In: Carette, J., et al. (eds.) Proceedings of Intelligent Computer Mathematics: MKM, Calculemus, DML, and Systems and Projects 2013 (CICM 2013). Lecture Notes in Computer Science, vol. 7961, pp. 152–167. Springer (2013). doi:10.1007/978-3-642-39320-4_10
14. Kaliszyk, C., Urban, J., Vyskočil, J., and Geuvers, H.: Developing Corpus-Based Translation Methods between Informal and Formal Mathematics: Project Description. In: Watt S. M., et al. (eds.) Proceedings of the International Conference on Intelligent Computer Mathematics (CICM 2014). Lecture Notes in Computer Science, vol. 8543, pp. 435-439. Springer, Cham (2014). doi:10.1007/978-3-319-08434-3_34
15. Wang, Q., Kaliszyk, C., and Urban, J.: First Experiments with Neural Translation of Informal to Formal Mathematics. In: Rabe, F., et al. (eds.) Proceedings of the 11th International Conference on Intelligent Computer Mathematics (CICM 2018). Lecture Notes in Computer Science, vol. 11006, pp. 255-270. Springer, Cham (2018). doi:10.1007/978-3-319-96812-4_22
16. Kaliszyk, C., Urban, J., and Vyskočil, J.: Learning to Parse on Aligned Corpora (Rough Diamond). In: Urban, C. and Zhang, X. (eds.) Proceedings of the 6th International Conference on Interactive Theorem Proving (ITP 2015). Lecture Notes in Computer Science, vol. 9236, pp. 227-233. Springer, Cham (2015). doi:10.1007/978-3-319-22102-1_15
17. Kaliszyk, C., Urban, J., and Vyskočil, J.: Automating Formalization by Statistical and Semantic Parsing of Mathematics. In: Ayala-Rincón, M. and Muñoz, C. (eds.) Proceedings of the 8th International Conference on Interactive Theorem Proving (ITP 2017). Lecture Notes in Computer Science, vol. 10499, pp. 12-27. Springer, Cham (2017). doi:10.1007/978-3-319-66107-0_2
18. Atanasyan, L., Butuzov, V., and Kadomcev S.: Geometry, 7-9 grades: textbook for general-education schools. Prosveshenie, Moscow (2010)