Evaluation of the difficulty of a geometric statement: comparing ChatGPT and GeoGebra Discovery

Piedad Tolmos
Department of Economía Financiera y Contabilidad, University Rey Juan Carlos
Madrid, Spain, 28032
piedad.tolmos@urjc.es

Tomas Recio and M.Pilar Vélez
Department of Matemáticas y Física, Universidad Antonio de Nebrija
Madrid, Spain,28015
trecio,pvelez@nebrija.es

Abstract

Our communication will present some initial results from an experience that we are developing, comparing the "complexity" measure assigned by GeoGebra Discovery's ShowProof command to a variety of well-known, elementary geometric statements, and the performance (i.e. correctness, clarity, and level of detail in the answer) of ChatGPT when asked about the same statements. Let us recall that the ShowProof command algorithmically outputs a proof by contradiction of a given geometric statement, expressing 1 as a combination of the hypotheses and the negation of the thesis. And ranks the "interest" or "difficulty" of the statement by computing the highest degree of the polynomials required to describe such contradiction. Measuring the interest of the output of automated reasoning tools is a classical challenge, but we think that the rank computed by the ShowProof command could be the first algorithmic approach towards establishing such measure in the context of geometric statements, although yet requiring a careful experimental work, such as the one we are initiating now, regarding its practical performance..

1 Introduction

GeoGebra Discovery[1] is a free, fork version of GeoGebra[2], with enhanced features concerning the automated verification and discovery of euclidean geometry statements. See references [5, 7, 2, 8], where these reasoning tools, as well as their development along the years and their possible educational applications, are described in detail.

Roughly speaking, concerning "verification", GeoGebra Discovery, through the *Prove* and *ProveDetails* commands, confirms or denies the truth of a statement (the latter, provides as well information

[1] https://kovzol.github.io/geogebra-discovery
[2] https://www.geogebra.org

about degenerate cases that should be avoided in the statement, such as the coincidence of the three vertices of a triangle). Regarding "discovery", the *Relation(p,q)* command, involving two geometric objects (e.g. lines) or two algebraic expressions about segment lengths (e.g. the perimeter and the area of a triangle), finds properties of the objects (e.g. the lines p, q are perpendicular) or of the expressions (e.g. if $a+b+c$ is the perimeter and $h \cdot c/2$ is the area, then *Relation*$(a+b+c, h \cdot c/2)$ outputs the inequality $(a+b+c)^2 \geq 6 \cdot \sqrt{(}3) \cdot (h \cdot c))$, that have not being previously formulated by the user.

This "discovery" ability is extended with the *Discover* command (or with the *Automated Geometer* web page[3], that outputs all properties of a certain kind (equality of distance between points, collinearity of three points, concyclicity of four points, parallelism or perpendicularity of segments defined by two points) that hold involving one given point in a figure (or all the elements of a figure, respectively).

Let us remark that in all cases this is accomplished by internally searching and/or verifying geometric properties through symbolic computations performed by Computer Algebra programs embedded in GeoGebra Discovery, therefore providing results that are mathematically correct (see [4]). And, on the other hand, we highlight the diversity of operating systems and devices (smartphones, laptops, tablets) where all these automated reasoning and discovery can be experimented by the user. Both facts: mathematical rigor, and ample availability, are specially relevant for considering the potential use of GeoGebra Discovery in the educational context.

2 The ShowProof command

In our contribution we focus on a particular issue (the algorithmic measure of the "complexity" of a geometric statement) regarding the performance of a recent feature of GeoGebra Discovery, the *ShowProof* command (see [6]). *ShowProof* begins displaying, for a given statement, each of the algebraic equations corresponding to the construction steps of the figure describing the statement. Finally, *ShowProof* algorithmically finds (using G-Basis computations) the expression of the *NormalForm* of 1 (i.e. 0) as a combination of the equations of hypotheses $h_1, \ldots, h_r$ and the negation of the thesis t, that is: $1 = g_1 \cdot h_1 + \cdots + g_r \cdot h_r + g_{r+1} \cdot (z \cdot t - 1)$, a proof by contradiction of the geometric statement. And ranks its "interest", "complexity" or "difficulty" by computing the highest degree of the polynomials $g_1, \ldots, g_{r+1}$.

This is a sort of "intrinsic" measure of the complexity of the interaction "thesis/ hypotheses", that we consider of particular interest in our context for two different reasons. One, because, as we have already described, GeoGebra Discovery has different ways of automatically providing a large number (thousands, some times) of properties holding on a figure, and a way to improve the performance and usefulness of GeoGebra Discovery could be to limit the obtained results to those within a certain rank of interest/difficulty, depending on and controlled by the expertise of the user. Two, because such filter could be quite useful in the educational context, for example, as a helpful tool concerning the formulation of adequate problems for math contests (see [1] for a related approach).

Measuring the interest of the output of automated reasoning tools is a classical challenge. See [9], where the author asks "to identify appropriate properties to permit an automated reasoning program to find new and interesting theorems, in contrast to proving conjectured theorems". But we

[3] `https://autgeo.online/ag/automated-geometer.html`

think that our proposal could the first algorithmic approach towards such measure, in the context of geometric statements, requiring a careful experimental work regarding its practical performance.

3 Our contribution

The analysis of the behavior of our proposed evaluation protocol through the *ShowProof* command is an on-going project, and our communication intends to present some initial results. Thus, in the context of the recent work [3] of our group concerning the possible cooperation of symbolic vs generative AI, i.e. GeoGebra Discovery/ChatGPT, that can be considered as the general framework of our contribution, we have developed an experience comparing the "complexity" measure assigned by *ShowProof* to a variety of well-known, elementary geometric statements, and the performance (i.e. correctness, clarity, and detailed answer), regarding the same statements, of ChatGPT.

We consider that the result of such comparison will allow us to establish some conclusions regarding the relation between the GeoGebra Discovery rank and ChatGPT reply readability and success for the same question, as a relevant step to advance towards the improvement of both softwares through their interaction. Finally, let us remark that the comparison of the obtained data (GeoGebra Discovery/ChatGPT) with respect to human (i.e. students or teachers of different educational levels) subjective perception of the interest or difficulty of a geometric statement, is also an on-going experience that is currently being developed by a group of closely related colleagues, associated members of the same research grant team *IAxEM-CM*.

4 Acknowledgement

Authors are partially supported by the project "Inteligencia aumentada en educación matemática mediante modelización, razonamiento automático e inteligencia artificial (IAxEM-CM)" (PHS-2024/PH-HUM-383), granted by the Comunidad de Madrid-Consejería de Educación, Ciencia y Universidades-Programas de Actividades de I+D entre Grupos de Investigación de la Comunidad de Madrid, Convocatoria Procesos Humanos y Sociales.

References

[1] Patrick Bak. Automated generation of planar geometry Olympiad problems, Ph.D. thesis, Master Thesis. Pavol Josef Safárik University, Kosice, (2020). `https://github.com/PatrikBak/GeoGen`

[2] Francisco Botana, Zoltán Kovács, Tomás Recio. A Mechanical Geometer, *Math.Comput.Sci.* 15, pp. 631–641, (2021). `https://doi.org/10.1007/s11786-020-00497-7`.

[3] Francisco Botana, Tomás Recio, M.Pilar Vélez. On Using GeoGebra and ChatGPT for Geometric Discovery, *Computers* 3(8), 187, (2024). `https://doi.org/10.3390/computers13080187`.

[4] Pierre Boutry, Gabriel Braun, Julien Narboux. Formalization of the arithmetization of Euclidean plane geometry and applications, *Journal of Symbolic Computation* 90, pp. 149–168, (2019.) `https://doi.org/10.1016/j.jsc.2018.04.007`.

[5] Zoltan Kovács, Bernard Parisse, Tomás Recio, M. Pilar Vélez. Automated reasoning tools in GeoGebra Discovery, *ACM Communications in Computer Algebra* 55 (2), pp. 39–43 (October 2021). https://doi.org/10.1145/3493492.3493495.

[6] Zoltán Kovács, Bernard Parisse, Tomás Recio, M. Pilar Vélez, Jonathan H. Yu. Towards the automated ranking of elementary geometry theorems, *ACM Communications in Computer Algebra* 58 (2), pp. 27–30 , (June 2024). https://doi.org/10.1145/3712023.3712026.

[7] Zoltán Kovács, Tomás Recio, M. Pilar Vélez. Automated reasoning tools with GeoGebra: What are they? What are they good for?, In Richard, P. R., Vélez, M. P., van Vaerenbergh, S., editors, *Mathematics Education in the Age of Artificial Intelligence: How Artificial Intelligence can serve mathematical human learning.* Mathematics Education in the Digital Era, 17, pp. 23–44. Springer Cham, (2022).

[8] Zoltán Kovács, Tomás Recio, M. Pilar Vélez. GeoGebra Automated Reasoning Tools: why and how (to use them in the classroom), *The 28th International Conference on Applications of Computer Algebra (ACA 2023). Program and Abstracts.* Warsaw University of Life Sciences SGGW. Institute of Information Technology. July 17–21, pp. 47–49, (2023). https://iit.sggw.edu.pl/wp-content/uploads/sites/18/2023/07/ACA2023_Program_Abstracts-1.pdf?x.

[9] Larry Wos. The problem of automated theorem finding, *Journal of Automated Reasoning* 10, pp. 137–138, (1993). https://doi.org/10.1007/BF00881868.