Reconstructing Materials Tetrahedron: Challenges in Materials Information Extraction

Kausik Hira[1], **Mohd Zaki**[2], **Dhruvil Seth**[1],
Mausam[1], **N. M. Anoop Krishnan**[1,2]
[1]Yardi School of Artificial Intelligence, [2]Department of Civil Engineering
Indian Institute of Technology Delhi
{kausikhira, mohdzaki1995, dhruvilsheth01}@gmail.com
{mausam, krishnan}@iitd.ac.in

Abstract

Discovery of new materials has a documented history of propelling human progress for centuries and more. The behaviour of a material is a function of its composition, structure, and properties, which further depend on its processing and testing conditions. Recent developments in deep learning and natural language processing have enabled information extraction at scale from published literature such as peer-reviewed publications, books, and patents. However, this information is spread in multiple formats, such as tables, text, and images, and with little or no uniformity in reporting style giving rise to several machine learning challenges. Here, we discuss, quantify, and document these outstanding challenges in automated information extraction (IE) from materials science literature towards the creation of a large materials science knowledge base. Specifically, we focus on IE from text and tables and outline several challenges with examples. We hope the present work inspires researchers to address the challenges in a coherent fashion, providing to fillip to IE for the materials knowledge base.

1 Introduction

Understanding a material's behaviour requires knowledge about its composition, properties, processing and testing protocols, and microstructure. The materials science (MatSci) tetrahedron shown in Fig. 1 encompasses the different aspects of a material which are reported by researchers in research papers, patents, and scientific documents. There exist wide variations in the reporting style of information in research papers making the automated MatSci IE a challenging task. Although researchers have developed rule-based systems, machine learning algorithms, and language models to represent and extract information from text, tables, and figures in MatSci research papers (see [21, 48] for detailed reviews), there exists no consolidated work which lists the outstanding challenges in MatSci IE and guides researchers to areas requiring more attention.

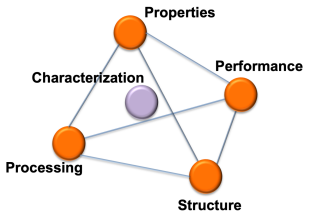

Figure 1: The MatSci tetrahedron

In this work, we thoroughly review MatSci articles and identify challenges related to the extraction of the aforementioned aspects of a material, as shown in Fig. 1. We also highlight the outstanding challenges of autonomously connecting extracted information together to create a large MatSci KB. Note that there exist millions of scientific documents reporting information about various materials known to humans; thus, automation in MatSci IE will lead to a rich knowledge-base on materials. The outline of the paper is as follows: First, we explain the methodology of collecting papers for review and annotation process. Then, in the results and discussion sections, we start by taking one component from the MatSci tetrahedron, report their locations in the articles and discuss the challenges in their extraction. We also quantify how frequently a challenge occurs to

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

motivate researchers in gauging the amount of information that will be obtained after solving respective challenges. Specifically, we identify challenges in extracting and connecting the information from text and tables present in MatSci research papers. Finally, we give some guidelines on information reporting protocols which are both machine and human-friendly to understand and that researchers can adopt to streamline automated MatSci IE from research papers. We then conclude the paper by indicating future research directions and the effect on the domain if the discussed outstanding challenges are solved.

2 Methodology

Recently in [11], a framework was presented to extract compositions from tables in MatSci articles, where the training data was annotated using distant supervision while the val and test data were manually annotated. We have searched for challenges in extracting each component of the MatSci tetrahedron from the enitre dataset (2536 articles), and then recorded the frequency of each challenge by randomly taking 50/100 tables containing the relevant information from the val and test dataset, or manually annotated 50 papers selected arbitarily as per requirement. All statistical figures reported in this paper rely on manual annotation either from [11] or done by us to ensure their accuracy. Further, we manually analyzed in a given document where the information is available, in tables or text. All the results and data associated with the annotation process are shared in the following link.

3 Results and Discussions

To highlight the outstanding challenges in IE related to each node of the MatSci tetrahedron, we randomly sampled 50 publications and identified where they are reported inside the research papers. Table 1 shows the the percentages of papers reporting raw materials (precursors), compositions, properties, processing, and testing methods in text and tables. Note that the same information could be reported in both text and table and hence the percentages may add to more than 100. Although we found that compositions are documented in both text and tables in most papers, an in-depth analysis revealed that only 33.21% of the total compositions were reported in the text, whereas 85.92% of compositions were present in tables. The overlap exists due to researchers mentioning the same composition in both text and tables. Processing and testing conditions are mostly reported in text, while in 80% articles, precursors are in text. 82% articles report properties in tables. In the following sections, we discuss the outstanding challenges that need to be addressed for completing the MatSci tetrahedron for a given material.

Table 1: Information in MatSci papers

	Text	**Table**
Precursors	80	10
Compositions	78	74
Properties	32	82
Processing	100	10
Testing	96	20

3.1 Composition extraction

Since the majority of the material compositions are reported in tables, we first discuss the outstanding challenges in extracting compositions from tables, followed by their extraction from text.

3.1.1 From tables:

We investigated 100 randomly selected composition tables from the manually annotated val-test data of [11] to report our empirical findings on each challenge's occurrence frequency.

a. Variation in table structure and information content: Based on table classification rules proposed by [11], tables can be of two types - Multi-Cell Composition (MCC) and Single-Cell Composition (SCC), which are sub-divided into tables containing complete information (CI) and partial information (PI), described further in 6.2.1 with Fig .7. The most prevalent were MCC-CI (36%), followed by SCC-CI (30%). PI tables are less common, with 24% being present as MCC-PI and the remaining 10% as SCC-PI, which is the outstanding challenge, remaining already addressed by [11] to a satisfactory level.

	Batched compositions (mol%)			Analyzed compositions (mol%)		
	SiO_2	B_2O_3	P_2O_5	SiO_2	B_2O_3	P_2O_5
XOI	55	25	20	56.0	25.8	18.2
XNG	55	22.5	22.5	63.4	26.9	9.7

Figure 2: Nominal (batch) and Analyzed Composition

b. Presence of nominal and experimental compositions: In 3% of the tables, we observed both nominal and analyzed/experimental compositions are reported. The nominal composition is the amount of chemicals taken initially to prepare the material, and analyzed/experimental composition refers to the actual composition of material obtained after analysing the manufactured material. (see Fig.2) [46] [32]. These values are not reported in any fixed pattern, which makes it difficult to separate the nominal and analyzed composition correctly after extraction.

c. Composition in table headers: Most tabular IE models like Tabbie, [15] DiSCoMaT, [11] perform better when row/column headers contain appropriate heading of its contents. In MatSci tables, the headers are mostly material IDs, compound names, properties, processing and testing labels, and the

inner cells contain corresponding values. However, in 6% of the tables, we found that the compounds with their values were present in the heading, which makes it hard for the IE models to extract the desired information. For instance - Se58Ge33Pb9 [16] or x=10%, x=20%,... [24] are column headers which contain both the compounds and corresponding concentration in the heading. 67% of these were SCC-CI, whereas the rest 33% were MCC-PI tables.

d. Compositions and related info inferred from other documents : In 11% of the tables [4] [5], we found that material composition and associated information are present in other research papers and given as references in the tables or their captions (see Fig. 3). Extracting the information from other papers will be challenging due to accessibility to those works.

Glass	σ_e^p (10^{-20}cm^2)	FWHM (nm)
Bismuth [18]	0.70	79
Tellurite [16]	0.75	65
Silicate [12]	0.55	40

Figure 3: References to other papers regarding composition.

3.1.2 From text:

Now, we report the challenges in extracting the compositions reported in the text of the MatSci research papers by investigating 50 randomly selected compositions from the manually annotated val-test data of [11] and reported our anecdotal findings on each challenge's occurrence frequency.

a. Different formats of compositions: The composition does not adhere to a predetermined pattern and encompasses various variations. Some notable examples are added to the appendix 6.2.2(a).

b. Extracting variable values in text: Extracting values from variables is challenging since the variable values are specified in different formats with some present only in text. For instance, "*A series of tellurite glasses with nominal composition (80-x)TeO_2–xGeO_2–10Nb_2O_5–10K_2O, where x = 0, 10, 20, 30, 40, 50, 60, 70 and 80 mol%, hereafter named 8T0G, 7T1G, 6T2G, 5T3G, 4T4G, 3T5G, 2T6G, 1T7G and 0T8G, respectively, were doped with 0.2 to 4 mol% $ErO_{1.5}$*" [25] has x values present only in text. A few other formats are added to the appendix 6.2.2(b).

c. Low recall in extracting compositions expressed with variables: 28% of the articles had compositions written with variables, within which 28.57% did not provide any values for the variables in the text. Among the 71.53% where values were present, 40% of them didn't mention the step size for the range of values taken by the variable, such as - "*x[0.75AgI: 0.25AgCl]: (1-x)[Ag_2O: WO_3], where 0.1≤x≤1 in molar weight fraction.*" [1] The step size of 0.1 was mentioned nowhere in the text but could be inferred from the composition table present in the paper. Therefore, extracting only from the text will cause a low recall, and this can be resolved by connecting the variables to the correct composition table containing the variable.

3.1.3 From table and text jointly:

Extracting information from PI tables is harder than extracting from CI tables, as we have to connect the table's partially mentioned composition correctly to one among many compositions mentioned in the text. We went through 50 manually detected PI tables noted in the val test data of [11] and found 36% of the tables to be far from the regular ones. We discussed some of the challenges and our statistical findings in this section and also demonstrated a regular MCC-PI table in Fig 8.

a. Unusual variables used: Other than the common variables x,X,y,z,Z, we also encounter variables like R, A, Y, and S in 4% of the manuscripts. Distinguishing some of them such as S or Y are difficult as they are valid symbols for elements [13].

b. Variables representing composition in text not found in tables: A generic way of extracting the composition when an arithmetic equation of the composition containing variables is mentioned in the text is to connect it to the variables located in the headings of the table and substitute it with values mentioned under it. Extraction becomes difficult if there is an absence or mismatch between the variable name in the table and the text, as shown in Fig. 4(a). We confronted 8% of the tables to have this challenge [45].

Table 2. Water content in xNa_2O-(100−x)GeO_2 glass

Glass	Water content[a] (moll^{-1})
0 Na_2O	0.016±0.008
5 Na_2O	0.022±0.006
10 Na_2O	0.019±0.003
15 Na_2O	0.021±0.005

(a)

Composition	x (mol%)	Density (g/cm^3)
60$GaS_{3/2}$ · 10GeS_2 · (30−x)$LaS_{3/2}$ · x$NdS_{3/2}$	0.3, 3.0	3.90±0.01
60$GaS_{3/2}$ · 10GeS_2 · (30−x)$LaS_{3/2}$ · x$HoS_{3/2}$	0.3, 3.0	3.92±0.01

(b)

Figure 4: (a) Variable **'x'** not in table. (b) Composition across multiple columns.

c. Composition present across multiple columns: The composition of the material is spread across multiple columns/rows (instance depicted in Fig. 4(b) [18]), or the table doesn't follow any fixed orientation; which makes retrieval extremely challenging. Seen among 4% of the PI tables.

3.2 Property extraction from tables

We detected the property tables and looked for challenges in their [11] entire dataset, and analyzed 100 arbitrarily selected property tables from their val_test dataset to state our empirical statistics.

a. Semantically similar row/column headers: We saw 19% of the tables having either headers with similar descriptions for separate entities or the abbreviations used to describe different properties are semantically very close (see Fig. 10). For example - heading of columns are Re2O3 density (g/cm^3) and Glass density (g/cm^3) [22], or in another table are T_g, T_{x1}, T_{x2}, T_{x3}, ΔT_x, T_m [10]. Identifying the desired property by a predictor model or someone without domain knowledge can be difficult.

b. Same property measured under different conditions: The same property can be measured with different techniques under different conditions. Therefore, it is important to extract the correct contextual information related to the reported property. Some recurrent scenarios include witnessing tables with various refractive indexes (RIs) at different wavelengths [33] (see Fig. 5), glass transition temperatures at different heating rates [9], or hardness at different testing loads. We encountered 9% of the property tables exhibiting this challenge.

c. Information in caption/footer instead of tables: Often, properties are mentioned with abbreviations in the headings of tables, which are semantically close to other properties (for example - Figure 10). The information regarding their abbreviation is commonly found in the caption or footer of the table. We observed 30% of the tables having this characteristic [39] [7]. We also saw 2% of the tables having no information on the properties' units but were found in the caption and footer of the tables [30]. Hence, text from these locations should also be considered for IE.

d. Multiple ways of reporting same unit: For instance, for g/cm^3 we found articles with gm/cm^3, g cm^{-3}, g/cm^3, gcm^{-3}, g/cc, gm/cc, gw/cm^3, gm cc^{-1}; and for kg/m^3, we witnessed kgm^{-3}, kg/m^3, kg m^{-3}. Extracting the correct unit and normalizing it to a standard form is essential for storing information on properties. Interestingly, while there are standard rules for writing SI units [42], as illustrated here, it is observed that these are not followed strictly in scientific publications.

3.3 IE or manufacturing and characterizing materials

a. Precursor extraction: A research paper generally investigates materials of a similar kind. Therefore, researchers rarely mention any material identifier in the text, hence it has to be assumed that all the materials are manufactured using the same precursors. In research papers where batch composition is mentioned in tables, the challenges are similar as mentioned in Section 3.1.1(b). In papers where researchers discuss the patented materials, they refer to them by their trademark name, for example, ***Pyrex, BOROSIL, Gorilla,*** etc. and hence their precursor information is not provided. However, papers discussing materials reported in previous publications, provide references to those papers reporting the required information in detail.

b. Processing conditions extraction: Let's consider the set of sentences describing the processing conditions: *"... powders were weighed and mixed thoroughly before being transferred to a **90 Pt/10 Rh crucible**, **heated** at **320°C** and **maintained** between **1000 and 1400°C** depending on composition, for approximately **25 min**. After **annealing** for approximately **three hours**, the glass was allowed to **cool slowly to room temperature**..."*. Hence, the challenges here are to extract temperatures and duration for each process, like heating, annealing, and cooling, along with the environmental conditions and experimental apparatus. Sometimes, these conditions are also mentioned in the table (see Fig. 5), and their extraction poses similar challenges as described in Section 3.2(b).

Table 1. Heat treated mode and some physical properties of the samples

Sample name	G	GC1	GC2	GC3	GC4
Heat treated time at 660°C	0h	4h	8h	16h	32h
Er^{3+} concentration (mol/dm^3)	1.796	1.785	1.783	1.776	1.784
Density (g/cm^3)	3.046	3.030	3.028	3.015	3.030
Thickness (cm)	0.150	0.156	0.156	0.156	0.156
n_C (at 656.3 nm)	1.527	1.530	1.530	1.530	1.525
n_D (at 589.3 nm)	1.530	1.533	1.532	1.532	1.528
n_F (at 486.1 nm)	1.535	1.539	1.538	1.538	1.533

Figure 5: Multiple challenges.

c. Testing conditions extraction: Consider the following example: *"The porous microstructure of the matrix was investigated by scanning electron microscopy **(SEM) (JEOL JSM T330A)**, by **infrared spectroscopy (IR)** in a **FT-IR spectrometer (Perkin Elmer Spectrum 2000)**, and by **X-ray powder diffractometry (XRD) (Siemens D-5000)**. The phase separation process was investigated by Raman microscope. The **room temperature Raman measurements** were performed through **Raman imaging microscope (Renishaw) system 3000**, with the **632.8 nm He–Ne laser** line for excitation"*. The testing conditions mainly comprise the sample characteristics, dimensions, test name, instrument name, instrument settings, and testing variables like temperature, wavelength, load, etc. The boldface text indicates the information to be extracted for obtaining a complete understanding of the testing process of a material. Fig. 5 lists different wavelengths at which a material is tested to

obtain wavelengths. The challenges faced in IE for this case will be similar to Section 3.2(b).
Material structure: To study the structure of materials, researchers perform x-ray diffraction studies, obtain the Raman spectra, optical micrographs, and scanning electron micrographs depending upon the depth of detail about the material structure required. This information is mostly reported in the figure and the figure description in the text provides some important details about the material's structure. In the statement, "*The **Raman spectrum** of the **porous** phase (Fig. 6(b)) shows only one **band at 277 cm**$^{-1}$ assigned to **silica** vibrations...*", the information about Raman spectra is already shown in the graph, and the text mentions only critical findings.

To summarise, the extraction of precursors, processing, and testing conditions from text poses challenges related to named entity recognition and relation extraction, which requires the need for specialized datasets and model development. This will also overcome the challenge of linking the extracted entities with the target materials. The challenges faced during IE from tables for these variables require overcoming similar challenges as explained earlier for composition (Section 3.1.1) and properties (Section 3.2)

3.4 MatSci Knowledge-base: Linking extracted information

The tetrahedron, as shown in Fig. 1, will be considered complete for a given material if its properties, processing, testing conditions, and raw materials required to manufacture are available. To this end, researchers need to connect extracted compositions with these variables, which can be done by connecting different paragraphs of the paper, text with tables, or tables with other tables in the paper, mainly through their material IDs. For instance, in [41], we obtain the composition of CAS1 from Table 1 and T_g of this material from Table 3. We detected 187 out of 2536 (7.37%) publications where inter-table IE is necessary and found difficulties in 81 of them.
a. Different material IDs in different tables: The same materials have been reported with different IDs in different tables. 21 out of 81 research papers (25.93%) have this challenge [35] [23].
b. Material IDs absent from tables: We detected a few MatSci papers where no IDs are present in the tables. There exist 23 out of those 81 documents (28.40%) having this challenge [31].
c. One of the tables doesn't contain material IDs: While connecting two tables; we found cases where IDs are mentioned only in one table [2] (37 out of 81 papers (45.67%) with this challenge). As the material ID is a very important factor in connecting tables, we did an intensive analysis on the type of IDs that are reported in the tables mentioned in the Table. 2 of the Appendix.

4 Guidelines for writing IE-friendly MatSci tables

Tables should be reported in such a way that automated extraction and the detection of the desired information are easy. Some of our suggestions are as follows:
a. Use MCC-CI tables with column orientation. (93% of MatSci tables follow column orientation).
b. The headers should contain the chemical formula of the constituent compounds/elements, along with the reported properties' acronyms, processing, and testing conditions, if any, written in a standard form. If precursors, processing, and testing conditions are common, they can be omitted from tables.
c. Units should be reported following the standard rules [42] in the column headers of the tables.
d. Material IDs should mandatorily be reported in the first column and formed as an acronym of its comprising constituents. There should be only one unique ID referring to a material in the article.
e. Table structure: [[ID], [C1], [C2], [P1], [P2], ...], where 'C' denotes the constituting compound/elements noted by their descending proportions, and 'P' refers to the properties of the material, should be followed.

5 Conclusion and future work

The literature is replete with IE challenges and algorithms to extract information about materials. However, there exists no study that quantifies how much benefit can be obtained if a particular challenge is solved. In this paper, we have identified and quantified several unresolved challenges present in IE for every aspect of the MatSci tetrahedron. Specifically, we pointed out the locations in a MatSci research paper, where each piece of information on the MatSci tetrahedron of a given material is reported. Further, we outline the challenges associated with the extraction of this information and linking them to build the MatSci KB. This will motivate researchers to focus on the outstanding challenges in the field, giving a direction on where to look and how much gain they might expect from resolving each challenge. In the end, we have suggested an IE-friendly table format, which will enhance the automated extraction of the desired information and improve the tabular understanding of the researchers. Such concerted efforts are required to streamline the reporting in MatSci articles, thereby accelerating IE towards materials discovery.

References

[1] RC Agrawal, ML Verma, and RK Gupta. Electrical and electrochemical properties of a new silver tungstate glass system: x [0.75 agi: 0.25 agcl]:(1- x)[ag2o: Wo3]. *Solid State Ionics*, 171 (3-4):199–205, 2004.

[2] M. Beggiora, I.M. Reaney, A.B. Seddon, D. Furniss, and S.A. Tikhomirova. Phase evolution in oxy-fluoride glass ceramics. *Journal of Non-Crystalline Solids*, 326-327:476–483, 2003. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(03)00456-3. URL `https://www.sciencedirect.com/science/article/pii/S0022309303004563`. 13th Int. Symp. on Non-Oxide Glasses and New Optical Materials.

[3] J.-L. Besson, G. Massouras, A. Bondanini, M. Huger, S. Hampshire, Y. Menke, and H. Lemercier. On the glass transition domain in some m-sialon (m = y or ln) oxynitride glasses. *Journal of Non-Crystalline Solids*, 278(1):187–193, 2000. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(00)00334-3. URL `https://www.sciencedirect.com/science/article/pii/S0022309300003343`.

[4] V.B. Brasil and E. Meyer. Test of the adiabatic nucleation model in sete and gesnse glasses. *Journal of Non-Crystalline Solids*, 348:7–10, 2004. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2004.08.118. URL `https://www.sciencedirect.com/science/article/pii/S002230930400660X`. Proceedings of the 6th Brazilian Symposium of Glases and Related Materials and 2nd International Symposium on Non-Crystalline Solids in Brazil.

[5] Qiuping Chen, Monica Ferraris, Daniel Milanese, Yvonne Menke, Elena Monchiero, and Guido Perrone. Novel er-doped pbo and b2o3 based glasses: investigation of quantum efficiency and non-radiative transition probability for 1.5 μm broadband emission fluorescence. *Journal of Non-Crystalline Solids*, 324(1-2):12–20, 2003.

[6] S De la Parra, LC Torres-Gonzalez, LM Torres-Martinez, and E Sanchez. Crystallization kinetics and phase transformation of xli2s–(1- x) sb2s3, x= 0–0.17 glass. *Journal of non-crystalline solids*, 329(1-3):104–107, 2003.

[7] O. Deparis, F.P. Mezzapesa, C. Corbari, P.G. Kazansky, and K. Sakaguchi. Origin and enhancement of the second-order non-linear optical susceptibility induced in bismuth borate glasses by thermal poling. *Journal of Non-Crystalline Solids*, 351(27):2166–2177, 2005. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2005.06.004. URL `https://www.sciencedirect.com/science/article/pii/S0022309305003959`.

[8] J.R. Duclère, A.A. Lipovskii, A.P. Mirgorodsky, Ph. Thomas, D.K. Tagantsev, and V.V. Zhurikhina. Kerr studies of several tellurite glasses. *Journal of Non-Crystalline Solids*, 355 (43):2195–2198, 2009. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2009.07.022. URL `https://www.sciencedirect.com/science/article/pii/S0022309309005109`.

[9] A Giridhar and Sudha Mahadevan. Mean atomic volume, tg and electrical conductivity of cux (as0. 4te0. 6) 100- x glasses. *Journal of non-crystalline solids*, 238(3):225–233, 1998.

[10] X Gu, LQ Xing, and TC Hufnagel. Glass-forming ability and crystallization of bulk metallic glass (hfxzr1- x) 52.5 cu17. 9ni14. 6al10ti5. *Journal of non-crystalline solids*, 311(1):77–82, 2002.

[11] Tanishq Gupta, Mohd Zaki, Devanshi Khatsuriya, Kausik Hira, N M Anoop Krishnan, and Mausam . DiSCoMaT: Distantly supervised composition extraction from tables in materials science articles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13465–13483, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.753. URL `https://aclanthology.org/2023.acl-long.753`.

[12] Maryam Habibi, Johannes Starlinger, and Ulf Leser. Deeptable: a permutation invariant neural network for table orientation classification. *Data Mining and Knowledge Discovery*, 34(6): 1963–1983, 2020.

[13] J. Henry and R.G. Hill. The influence of lithia content on the properties of fluorphlogopite glass-ceramics. i. nucleation and crystallisation behaviour. *Journal of Non-Crystalline Solids*, 319(1):1–12, 2003. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(02)01958-0. URL https://www.sciencedirect.com/science/article/pii/S0022309302019580.

[14] DB Hollis, FR Cruickshank, and MJP Payne. Structural influence of tellurite glasses on fluorescence of thulium near 1.86 μm. *Journal of non-crystalline solids*, 293:422–429, 2001.

[15] Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. Tabbie: Pretrained representations of tabular data, 2021.

[16] Praveen K Jain, KS Rathore, NS Saxena, et al. Structural characterization and phase transformation kinetics of se58ge42- xpbx (x= 9, 12) chalcogenide glasses. *Journal of non-crystalline solids*, 355(22-23):1274–1280, 2009.

[17] Yoann Jestin, A Le Sauze, B Boulard, Y Gao, and P Baniel. Viscosity matching of new pbf2–inf3–gaf3 based fluoride glasses and zblan for high na optical fiber. *Journal of non-crystalline solids*, 320(1-3):231–237, 2003.

[18] K. Kadono, M. Shojiya, M. Takahashi, H. Higuchi, and Y. Kawamoto. Radiative and non-radiative relaxation of rare-earth ions in ga2s3–ges2–la2s3 glasses. *Journal of Non-Crystalline Solids*, 259(1):39–44, 1999. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(99)00493-7. URL https://www.sciencedirect.com/science/article/pii/S0022309399004937.

[19] AH Khafagy, AA El-Adawy, AA Higazy, S El-Rabaie, and AS Eid. The glass transition temperature and infrared absorption spectra of:(70- x) teo2+ 15b2o3+ 15p2o5+ xli2o glasses. *Journal of non-crystalline solids*, 354(14):1460–1466, 2008.

[20] Tomoya Konishi, Tomoyoshi Asano, Yuu Ishii, Kohei Soga, Hiroyuki Inoue, Akio Makishima, and Satoru Inoue. Effects of eu2o3 on liquid–liquid phase separation of pbo–b2o3, bao–b2o3 and sro–b2o3 glasses. *Journal of Non-Crystalline Solids*, 265(1):19–28, 2000. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(99)00878-9. URL https://www.sciencedirect.com/science/article/pii/S0022309399008789.

[21] Olga Kononova, Tanjin He, Haoyan Huo, Amalie Trewartha, Elsa A Olivetti, and Gerbrand Ceder. Opportunities and challenges of text mining in materials research. *Iscience*, 24(3), 2021.

[22] J. Marchi, D.S. Morais, J. Schneider, J.C. Bressiani, and A.H.A. Bressiani. Characterization of rare earth aluminosilicate glasses. *Journal of Non-Crystalline Solids*, 351(10):863–868, 2005. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2005.01.078. URL https://www.sciencedirect.com/science/article/pii/S0022309305001080.

[23] J.R Martinelli, F.F Sene, and L Gomes. Synthesis and properties of niobium barium phosphate glasses. *Journal of Non-Crystalline Solids*, 263-264:263–270, 2000. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(99)00638-9. URL https://www.sciencedirect.com/science/article/pii/S0022309399006389.

[24] L. Minati, G. Speranza, M. Ferrari, Y. Jestin, and A. Chiasera. X-ray photoelectron spectroscopy of erbium-activated-silica–hafnia waveguides. *Journal of Non-Crystalline Solids*, 353(5):502–505, 2007. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2006.10.018. URL https://www.sciencedirect.com/science/article/pii/S002230930601372X. SiO2, Advanced Dielectrics and Related Devices 6.

[25] Gonçalo Monteiro, Luís F. Santos, Rui M. Almeida, and Francesco D'Acapito. Local structure around er3+ in geo2–teo2–nb2o5–k2o glasses and glass-ceramics. *Journal of Non-Crystalline Solids*, 377:129–136, 2013. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2012.12.019. URL https://www.sciencedirect.com/science/article/pii/S002230931200734X. ISNOG 2012 Proceedings of the 18th International Symposium on Non-Oxide and New Optical Glasses Rennes, France, July 1-5, 2012.

[26] M Mortier, P Goldner, P Féron, G.M Stephan, H Xu, and Z Cai. New fluoride glasses for laser applications. *Journal of Non-Crystalline Solids*, 326-327:505–509, 2003. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(03)00461-7. URL https://www.sciencedirect.com/science/article/pii/S0022309303004617. 13th Int. Symp. on Non-Oxide Glasses and New Optical Materials.

[27] Manoj Kumar Narayanan and HD Shashikala. Thermal and optical properties of bao–caf2–p2o5 glasses. *Journal of Non-Crystalline Solids*, 422:6–11, 2015.

[28] V. Nazabal, E. Fargin, B. Ferreira, G. Le Flem, B. Desbat, T. Buffeteau, M. Couzi, V. Rodriguez, S. Santran, L. Canioni, and L. Sarger. Thermally poled new borate glasses for second harmonic generation. *Journal of Non-Crystalline Solids*, 290(1):73–85, 2001. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(01)00726-8. URL https://www.sciencedirect.com/science/article/pii/S0022309301007268.

[29] Virginie Nazabal, S Todoroki, A Nukui, T Matsumoto, S Suehara, T Hondo, T Araki, S Inoue, Clara Rivero, and Thierry Cardinal. Oxyfluoride tellurite glasses doped by erbium: thermal analysis, structural organization and spectral properties. *Journal of non-crystalline solids*, 325 (1-3):85–102, 2003.

[30] Masayuki Okuno, Nikolay Zotov, Martin Schmücker, and Hartmut Schneider. Structure of sio2–al2o3 glasses: Combined x-ray diffraction, ir and raman studies. *Journal of Non-Crystalline Solids*, 351(12):1032–1038, 2005. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2005.01.014. URL https://www.sciencedirect.com/science/article/pii/S0022309305000050. Proceedings of the International Conference on Non-Crystalline Materials (CONCIM).

[31] O. Peitl and E.D. Zanotto. Thermal shock properties of chemically toughened borosilicate glass1the results described in this article were judged by an international committee and were considered to be the best poster of scientific nature presented at the iii brazilian symposium on glass and related materials, bonito, brazil, 1998.1. *Journal of Non-Crystalline Solids*, 247(1): 39–49, 1999. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(99)00029-0. URL https://www.sciencedirect.com/science/article/pii/S0022309399000290.

[32] K Petkov and P.J.S Ewen. Photoinduced changes in the linear and non-linear optical properties of chalcogenide glasses. *Journal of Non-Crystalline Solids*, 249(2):150–159, 1999. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(99)00330-0. URL https://www.sciencedirect.com/science/article/pii/S0022309399003300.

[33] Xvsheng Qiao, Xianping Fan, Jin Wang, and Minquan Wang. Luminescence behavior of er3+ ions in glass–ceramics containing caf2 nanocrystals. *Journal of Non-Crystalline Solids*, 351(5): 357–363, 2005. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2004.11.021. URL https://www.sciencedirect.com/science/article/pii/S0022309304011196.

[34] J Rocherullé, M Matecki, and Y Delugeard. Heat capacity measurements of mg–y–si–al–o–n glasses. *Journal of Non-Crystalline Solids*, 238(1):51–56, 1998. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(98)00578-X. URL https://www.sciencedirect.com/science/article/pii/S002230939800578X.

[35] M Romero, R.D Rawlings, and J.Ma Rincón. Crystal nucleation and growth in glasses from inorganic wastes from urban incineration. *Journal of Non-Crystalline Solids*, 271(1):106–118, 2000. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(00)00082-X. URL https://www.sciencedirect.com/science/article/pii/S002230930000082X.

[36] Rodrigo Santos, Luís F. Santos, and Rui M. Almeida. Optical and spectroscopic properties of er-doped niobium germanosilicate glasses and glass ceramics. *Journal of Non-Crystalline Solids*, 356(44):2677–2682, 2010. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2010.03.036. URL https://www.sciencedirect.com/science/article/pii/S0022309310002474. 12th International Conference on the Physics of Non-Crystalline Solids (PNCS 12).

[37] R. Sato and M. Tomozawa. Interaction of water diffusion and hydrogen diffusion in silica glass at high temperatures. *Journal of Non-Crystalline Solids*, 343(1):26–32, 2004. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2004.07.048. URL https://www.sciencedirect.com/science/article/pii/S0022309304005976.

[38] Cornelia Schröder, Mariana de Oliveira Carlos Villas-Boas, Francisco C. Serbena, Edgar D. Zanotto, and Hellmut Eckert. Monitoring crystallization in lithium silicate glass-ceramics using 7li→29si cross-polarization nmr. *Journal of Non-Crystalline Solids*, 405:163–169, 2014. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2014.09.008. URL https://www.sciencedirect.com/science/article/pii/S0022309314004633.

[39] S. Scudino, J. Eckert, C. Mickel, and L. Schultz. On the amorphous-to-quasicrystalline phase transformation in ball-milled and melt-spun zr58.5ti8.2cu14.2ni11.4al7.7 glassy alloys. *Journal of Non-Crystalline Solids*, 351(10):856–862, 2005. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2005.02.004. URL https://www.sciencedirect.com/science/article/pii/S002230930500116X.

[40] D Singh, S Kumar, and R Thangaraj. Optical and electrical properties of as-prepared and annealed (se80te20) 100 - xagx ($0 \leq x \leq 4$) ultra-thin films. *Journal of non-crystalline solids*, 358(20):2826–2834, 2012.

[41] M. Solvang, Y.-Z. Yue, S.L. Jensen, and D.B. Dingwell. Rheological and thermodynamic behavior of calcium aluminosilicate melts within the anorthite–wollastonite–gehlenite compatibility triangle. *Journal of Non-Crystalline Solids*, 351(6):499–507, 2005. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2005.01.034. URL https://www.sciencedirect.com/science/article/pii/S0022309305000487.

[42] Ambler Thompson and Barry N. Taylor. *Guide for the Use of the International System of Units (SI)*. National Institute of Standards and Technology, 2008.

[43] Koichiro Tsuzuku, Seiichi Taruta, Nobuo Takusagawa, and Hiroshi Kishi. Crystallization of 2(ca,sr,ba)o–tio2–2sio2 composition glasses. *Journal of Non-Crystalline Solids*, 306(1): 50–57, 2002. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(02)01057-8. URL https://www.sciencedirect.com/science/article/pii/S0022309302010578.

[44] G. Walter, U. Hoppe, A. Barz, R. Kranold, and D. Stachel. Intermediate range structure of mixed phosphate glasses by x-ray diffraction. *Journal of Non-Crystalline Solids*, 263-264: 48–60, 2000. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(99)00622-5. URL https://www.sciencedirect.com/science/article/pii/S0022309399006225.

[45] Satoshi Yoshida, Jun Matsuoka, and Naohiro Soga. Sub-critical crack growth in sodium germanate glasses. *Journal of Non-Crystalline Solids*, 316(1):28–34, 2003. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(02)01934-8. URL https://www.sciencedirect.com/science/article/pii/S0022309302019348.

[46] R.E Youngman, B.G Aitken, and J.E Dickinson. Multi-nuclear nmr studies of borosilicophosphate glasses and microfoams. *Journal of Non-Crystalline Solids*, 263-264:111–116, 2000. ISSN 0022-3093. doi: https://doi.org/10.1016/S0022-3093(99)00626-2. URL https://www.sciencedirect.com/science/article/pii/S0022309399006262.

[47] Maria Zaharescu, Alexandra Barau, Luminita Predoana, Mariuca Gartner, Mihai Anastasescu, Jan Mrazek, Ivan Kasik, and Vlastimil Matejec. Tio2–sio2 sol–gel hybrid films and their sensitivity to gaseous toluene. *Journal of non-crystalline solids*, 354(2-9):693–699, 2008.

[48] Mohd Zaki, Amreen Jan, NM Anoop Krishnan, and John C Mauro. Glassomics: An omics approach toward understanding glasses through modeling, simulations, and artificial intelligence. *MRS Bulletin*, pages 1–14, 2023.

[49] K. Zhao, Q. Luo, D.Q. Zhao, H.Y. Bai, M.X. Pan, and W.H. Wang. Bulk metallic glasses based on binary rare earth elements. *Journal of Non-Crystalline Solids*, 355(16):1001–1004, 2009. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2009.04.007. URL https://www.sciencedirect.com/science/article/pii/S0022309309001215.

6 Appendix

In the main paper, we discussed unsolved challenges while extracting each feature of the material tetrahedron from published research works. Here, in this section, we will address some more challenges besides those that were mentioned in the article. Some of them have been solved, which we will mention; while the rest remain unsolved.

The details of all the research papers used in this study, along with annotations to identify the challenges, are available at `https://anonymous.4open.science/r/MatSci-IE-Challanges-AI4MAT`

6.1 Common Challenges faced for during information extraction from tables

We begin by discussing the problems encountered for all-encompassing IE tasks. While challenge a has been resolved in [11] and b in [12] [11], no solution for challenge c was seen.

a. Distractor rows or columns: Additional contents in the table that are irrelevant to our desired information.

b. Different orientation of tables: Each table can have either of the two orientations - row or column, which is essential to recognize in order to extract information precisely. We saw 100 random composition tables and 100 random property tables and observed that 7% of the tables are represented with row orientation (see Fig.5), whereas 93% of the tables are represented with column orientation (see Fig. 8).

c. Multiple tables merged in one: A rarely seen challenge (<1%) illustrated in Figure 6, where many tables are concatenated in a long or broad table, which leads to difficulties in extracting the targeted details. [44]

Table 1. Compositions and densities for the mixed metaphosphate glasses studied

Mol% by analysis		Density (g/cm³)	Mol% by analysis		Density (g/cm³)	Mol% by analysis		Density (g/cm³)	Mol% by analysis		Density (g/cm³)
BaO	Na_2O		BaO	SrO		K_2O	Na_2O		CaO	MgO	
0.0	50.5	2.507	0.0	49.8	3.130	0.0	50.5	2.507	0.0	51.3	2.44
4.7	50.3	2.628	2.5	47.5	3.172	2.2	47.3	2.500	12.5	37.5	2.52
6.9	44.4	2.604	5.0	45.1	3.201	4.8	45.3	2.501	25.0	25.0	2.57

Figure 6: Multiple tables concatenated to form larger table.

6.2 Other challenges faced in composition extraction:

6.2.1 From tables:

We start by illustrating the different types of tables mentioned in Section 3.1.1(a) in Figure 7. When the entire composition is written within a single cell, it is classified as an SCC table [29][47], whereas when the composition is written across multiple cells of the table by reporting the value of each constituent (compounds or elements) of the composition in separate cells, it is defined as a MC table [17][27]. Now if the table contains all the constituents of the material, we can say it as CI table (complete information). But if it contains only some constituents, like mentioning the value of only one compound of the material, we then need to look into the text of the article and connect it with the table in-order to extract the composition completely. These tables are named as PI tables (partial information). Often MCC-PI tables are written using mathematical expressions with variables representing a particular constituent [19], as shown in Figure 8.

MCC-CI

Sample ID	P_2O_5	BaO	CaF_2
BaPF0	50	50	0
BaPF2	50	48	2
BaPF4	50	46	4

MCC-PI

Glass ID	Mol% PbF2	n_D
ZBLAYLP10	10.0	1.568
ZBLAYLP10	15.0	1.564

Table Footer text: ZBLAYLPx=56ZrF_4 (29−x)BaF_2–xPbF_2–3.5LaF_3–2YF_3–2.5AlF_3–7LiF.

SCC-CI

Composition	Tg ±5 °C
75TeO_2–25ZnO	315
75TeO_2–18ZnO–7ZnF_2	278
75TeO_2–9ZnO–16ZnF_2	261

SCC-PI

Composition	Sensitivity
5% TiO_2	0.22 ± 0.01
10% TiO_2	0.28 ± 0.01
15% TiO_2	0.19 ± 0.01

Figure 7: Classification of composition tables

We discuss six more challenges which can be seen in the composition tables. Challenge **a** has already been handled by [11]. In challenge **b**, extraction of compositions mentioned with atomic%, atomic fraction and parts per million (ppm) is still outstanding, whereas extraction of dopant concentration from challenge **c** has not been solved yet. The rest of the Challenges - **d, e, and f** haven't been addressed by the community yet and should be given attention too.

Table 2. Variation of relative intensity, I_r, of some of absorption bands with composition of the glasses: (70−x) TeO_2+15B_2O_3 +15P_2O_5+xLi_2O

X (mol%)	Harmonic P–O–P bending	TeO_3 tp and TeO_4 tbp	BO_4units+v_3 PO_4^{3-} ion	Ionic B–O stretching	Stretching BO_3 units
5	0.072	0.133	0.139	–	0.131
10	0.071	0.131	0.137	–	0.129
15	0.073	0.115	0.119	–	0.108

Figure 8: Typical MCC-PI table

a. One Composition with multiple units: Consider the following example composition - 0.85TeO_2+0.15WO_3+0.1wt%Ag_2O+0.076wt%CeO_2 [8]. Here, for a given material, different components are measured in different units (mol% and wt%). This is found in 2% of the tables making composition extraction challenging.

b. Composition expressed with different units in various articles: such as mol%, weight%, atomic%, mol fraction, weight fraction, atomic fraction, and ppm. Among them, the most commonly used unit is mol%, followed by weight%.

c. Percentage not equal to 100: In some papers, even after extracting the whole composition correctly, we observe that the sum of the coefficient doesn't sum up to 100, whereas we also notice the presence of the reverse scenario. Especially in the case of doping, the sum exceeds 100, which is correct. The challenge is to identify where we need to normalize the values extracted and where we shouldn't. We noted that dopant is reported in 2% of the composition tables.

d. Compounds with different units in the same table: Sometimes, the same composition is recorded with multiple units having different values, which makes it hard to separate duplicate data. We encountered this situation in 2% of the composition tables. [20]

e. Compounds inferred from material ID: In 7% of the tables, we confronted IDs from which valuable information regarding the composition was present. For example - some of them had the molar ratios of compounds present in the glass, while others had the composition written in an abbreviated form. We also found IDs where composition has been mentioned as a substring in it. [28]

f. Variables used to represent compounds: When a composition is expressed with variables, it mostly denotes the various concentrations of the compounds in different materials. However, in some articles, we came across compositions where variables have been used to represent compound names instead of their values. One such example is: RE36Y20Al24Co20 (RE = Ce, Pr, Nd, Sm, Gd, Tb, Er, Sc) [49]. We saw the presence of this scenario in 1% of the tables.

6.2.2 From text

a. Different formats of compositions examples: A few instances of such different formats are:
1. *"Erbium-doped glasses with the molar composition 40GeO_2.10SiO_2.25Nb_2O_5.25K_2O, plus 0.1 to 4 mol% of ErO1.5, were prepared using mixtures of the respective oxides (99.99% purity), with exception of K2O, which was added in the form of K2CO3"* [36].
2. *"Bulk samples of $(Se_{80}Te_{20})_{100-x}Ag_x$ ($0 \leq x \leq 4$) system were prepared by conventional melt quenching technique. High-purity (99.999%) elements with appropriate atomic percentages were sealed in a quartz ampoule (length $\sim$ 100 mm and internal diameter $\sim$ 6 mm) in a vacuum of 10 - 5 mbar"* [40].
3. *"The samples having chemical composition of 2(Ca,Sr,Ba)O–TiO_2–2SiO_2 were examined. CaO, SrO and BaO contents in the samples were varied as shown in Table 1. RO% shows molar percent of CaO, SrO or BaO in total RO of CaO+SrO+BaO."* [43]

b. Extracting variable values in text: Extracting values from variables is challenging since the variable values are specified in different formats with some present only in text. For instance, *A series of tellurite glasses with nominal composition (80-x)TeO_2–xGeO_2–10Nb_2O_5–10K_2O, where x = 0, 10, 20, 30, 40, 50, 60, 70 and 80 mol%, hereafter named 8T0G, 7T1G, 6T2G, 5T3G, 4T4G, 3T5G, 2T6G, 1T7G and 0T8G, respectively, were doped with 0.2 to 4 mol% $ErO_{1.5}$* [25] has x values present only in text. A few instances of other formats of composition with variables are shown below:

1. *"The non-isothermal crystallization kinetics of xLi_2S–(1-x)Sb_2S_3, x=0–0.17 were investigated using differential scanning calorimetry (DSC)"* [6].
2. *"To ascertain the effect of the glass composition on fluorescence parameters around 1.86 μm, we*

prepared and experimented on two series of glasses. The first one was $aR^1{}_2O(1\text{-}a)TeO_2$ where 'a' was 0, 10, 15, 20, 30 mol%, and 'R^1' was Li, Na, K. The second one was $bR^{11}O.cR_2{}^{111}O_3(1\text{-}b\text{-}c)TeO_2$ where 'b' was 0, 10, 20, 30 mol%, and 'c' was 0.5% or 16.5%, and 'R^{11}' = Ba, 'R^{111}' = Al, Ga, or In. To find the effect of concentration quenching, the concentration of thulium oxide was varied from 0.01 to 5.0 mol%" [14].

3. *"Glasses with composition in mol%: 51ZrF_4, 16BaF_2, 5LaF_3, 3AlF_3, 20LiF, 5PbF_2 have been prepared by melting of the powders (commercial raw materials of purity higher than 99.99%) in a covered vitreous carbon crucible at about 850 °C for 45 min in a dry argon glove box with a water content lower than 5 ppm. The melt was poured into a preheated copper mould at 240 °C and slowly cooled down to room temperature. The doping ion was added in excess to the formula +xErF_3 from 0.01 to 11 mol% corresponding to 0.02 to 22 × 1020 Er3+ ions/cm3. The samples obtained were of good optical quality*" [26].

c. Unit not mentioned: 39.53% compositions had no unit specified explicitly.

d. Percentages not summing to 100: Out of 78% compositions found in the text, 17.94% of them did not have the sum of coefficients of the compounds equal to 100 and needed to be normalized.

e. Recognition of full forms and abbreviations: Instead of providing precise composition values, full forms are employed in lieu of abbreviations. A few instances are:

1. *"Lithium disilicate glass was prepared in 30 g quantity by heating stoichiometric homogeneous mixtures of lithium carbonate (99.0%), Synth, and silica (99.9999%), Santa Rosa, for 4 h at 1500 °C in a platinum crucible.* has the composition of lithium disilicate indirectly." [38]

2. *"The silica glass samples used in the present experiment were prepared by melting quartz crystals at 1800 °C for 1 h in a one atmosphere hydrogen gas, and they contained nearly the same amount of SiOH and SiH* has the composition of silica (SiO_2) indirectly." [37]

f. Unstable composition extraction: Unstable reagents are also identified as the composition due to a lack of robust parsers. AlO_4 is used as a reagent while $SiO2$ can be a composition.

6.2.3 From table and text jointly:

As extraction from PI is harder than CI tables, it will benefit the community if a standard form is followed for the PI tables. Figure 8 is a good example of a standard MCC-PI tables from which we can extract the composition without much difficulties. Let us look into the less-seen PI challenges.

a. Composition partly in the table and partly in text: Although we know that PI tables contain the composition partly, we expect the text to contain the full composition. But in seldomly seen occurrences, as depicted by Figure 9, we found that only the remaining part of the composition, which isn't mentioned in the table, is present in the text, which makes connecting and extracting the whole composition extremely difficult. This was seen in less than 1% of the PI tables.

Table 3. Physical properties of glasses containing 5% AlF_3 and 10% BaF_2

ZrF_4	NaF	LiF	n_D±0.001	T_g±2 (°C)	T_x±2 (°C)	T_{p1}±2 (°C)
50	20	15	1.483	17	325	335
40	30	15	1.469	203	266	300
40	20	25	1.475	200	321	336

Figure 9: Some part of the composition in table, rest of it in the text.

6.3 Other challenges faced in property extraction from tables:

Before discussing the other unresolved challenges faced during property extraction, we would like to exhibit Fig.10 [10] in order to demonstrate the table seen in Section 3.2(a) and (b).

a. Same property reported with different units: Scarce occurrence (1%) of property tables were seen where two columns/rows of the same property having values in different units were recorded, which can lead to duplication of data while extracting. For instance - glass transition temperature (T_g) is mentioned in both °C and K [34]. This problem might be tackled by post-processing with rule-based systems.

Table 1. Composition dependence of the density and the characteristic temperatures of $(Hf_xZr_{1-x})_{52.5}Cu_{17.9}Ni_{14.6}Al_{10}Ti_5$ (x=0, 1/3, 1/2, 2/3, and 1), where T_{xi} represents the temperature of the ith exotherm, ΔT_x is the undercooling range, T_{rg} is the ratio of the glass transition temperature (T_g) over the melting temperature (T_m)

x	ρ (g/cm³)	T_g (K)	T_{x1} (K)	T_{x2} (K)	T_{x3} (K)	T_{x4} (K)	ΔT_x (K)	T_m (K)	T_{rg}
0	6.65	686	725	740	792	855	39	1075	0.638
1/3	8.16	708	753	766	798	928	45	1126	0.629

Figure 10: Property description in caption & semantically close headings.

b. Range of values (min-max) given instead of mean value: Seldomly (<1%), we encounter tables where the values of properties are mentioned in range rather than a fixed value. For example - we saw T_g values mentioned in

as 930-945 °C [3]. One has to make the decision of taking the min, max, or mean of the documented values.

c. Property recorded under various acronyms: It is a common practice to record property names with their abbreviations. Some properties can have various abbreviations, like density is represented with either ρ or d, Young's modulus with YM or E, activation energy with E0, Ae, Ea, or E a.

6.4 ID Analysis:

As material ID is the key component in connecting glasses from tables to text, or between two different tables, or also in between different sections of the text, we went through arbitrarily selected 25 articles containing material IDs in the tables and recorded their semantic pattern in order to observe the semantics used by authors to refer to their materials. We found that a whopping majority of the authors prefer to use acronyms or self-made codes to refer to the glasses, followed by natural numbers, alphabets, and standard glass names. Occasionally, we also see IDs containing relevant information like the processing conditions of the material, or information about the state of the glass like amorphous or crystalline, and its composition. ID are generally present at the beginning of the table, and very rarely we see an ID column to be present in the middle or end. Its not a very good idea to make IDs inter-dependable, which can cause difficulties in extracting the materials.

Challenges	**% of occurence**
Standard Glass Name	8
Acronyms/Codes	80
Alphabets	8
Natural Numbers	12
Decimal Numbers	ϵ
Acronym* + Processing Cond as ID	8
ID Present in Middle	ϵ
Apparently Two IDs Present	4
Ratios as IDs	ϵ
Composition and State of Glass Info in ID	4
Contains Info About Processing Condition	20
Same IDs but Different Composition	ϵ
IDs Interconnected	ϵ

Table 2: Various forms of ID encountered

Note: * represents may or may not be present, ϵ represents <1% in Table 2.