NLP in the Legal Profession: How about Small Languages?

Csaba Csáki[1], Péter Homoki[2], György Görög[3] and Pál Vadász[4]

[1]Corvinus University of Budapest, Fővám tér 8, 1093 Budapest, Hungary
[2]Homoki Law Office, Katona József u. 39, 1137 Budapest, Hungary
[3]Montana Ltd, Budapest, Hungary
[4]National University of Public Service, Ludovika tér 2, 1083 Budapest, Hungary

Abstract

Over the last three decades legal service providers as well as legal departments of various firms have embraced the opportunity to apply the latest digital technology to improve the efficiency and effectiveness of their work. Since language is central to both law-making and during the application of the law, Natural Language Processing solutions have found their way to this profession. One particular research area relates to the issue of small languages. The problem is rooted in the size of the population speaking a given language: in a small market, it is not economically feasible to develop NLP technologies as they require considerable time and effort to develop a sufficient language corpus. This paper reviews the challenges countries and jurisdictions of small languages face in light of increasing NLP applications in legal contexts, while also examining the role of the public sector in relation to addressing such issues.

Keywords

Natural Language Processing, Legal Tech, small languages, AI regulations

1. Introduction: NLP-based Legal Tech and Small Languages

The use of information and communication technologies (ICT) in the field of law is often referred to as legal technology (or Legal Tech, sometimes LegalTech for short) [1]. While the term - similarly to other terms related to the application of IT - has a few possible interpretations [2], more recently it alludes to the trend of using the latest technologies in the legal sector [3]. This lack of clear definition comes from the evolutionary nature of ICT where both the actual technology and its scope (goals and application) change fast and require quick adaptation. One of the most recent trends pushing the boundaries of Legal Tech is based on artificial intelligence (AI) and specifically Natural Language Processing (NLP) [4, 5].

The application of information technology to the legal profession is certainly not new [6]. However, over the last three decades, many different legal users (legal service providers, legal departments of various firms, and even judges and prosecutors) have embraced the opportunity to apply the latest digital technology to improve the efficiency and effectiveness of their work [3]. So much so, that a new field of legal technology has emerged including systems like

EGOV-CeDEM-ePart 2022, September 06–08, 2022, Linköping University, Sweden (Hybrid)
csaki.csaba@uni-corvinus.hu (C. Csáki); peter.homoki@homoki.net (P. Homoki); goroggy@gmail.com (G. Görög); pal.vadasz@uni-nke.hu (P. Vadász)
0000-0002-8245-1002 (C. Csáki); - (P. Homoki); - (G. Görög); 0000-0003-3848-6096 (P. Vadász)

legal document and contract management, digital or virtual data room, automated document assembly, legal-research analysis, legal-practice management, e-discovery, or judicial predictive systems [7, 8]. Since language is central to law-making as well as during the application of the law [1, 9] it is no wonder that NLP solutions have found their way to this profession. This is true even when considering the 'messiness' of law practice [10]. Natural Language Processing is a fast-growing segment of the AI field [9]. NLP is a computer-based approach to analysing text and speech that relies on a solid theoretical background as well as a range of technical solutions [11]. NLP in the legal sector is used for almost every aspect of the work and is utilized to some extent by almost every IT system mentioned above.

One particular research area of NLP application relates to the issue of small languages [12, 13]. The problem is rooted in the size of the market, that is the size of a population speaking a given language. Unlike in Ethnologue, where a language is considered small when it has less than 10,000 speakers, in this study the focus is the digital presence and NLP, therefore, 'small' is understood as a language with less than 10 million speakers. Such small languages are disadvantaged both from a technical and from an economic point of view. In a small market, it is not immediately economically feasible to develop NLP technologies as they require considerable time and effort to develop a sufficient language-specific base (the language corpus). Even with an acceptable starting corpus, domain or profession-specific frameworks (such as a legal corpus) would also be needed before the final development of a particular application or solution.

The purpose of this investigation, therefore, is to review the challenges countries and jurisdictions of small languages face in light of increasing NLP applications in legal contexts, while also examining the role of the public sector in relation to such issues. The paper first looks at the role of language in the legal profession, then reviews the role of NLP in various contexts of the legal profession. After introducing the small language problem, an overview of potential solutions (from practice and as they appear in the scientific literature) is provided with a special focus on the role of the public sector in solving the issue. A summary and conclusions complete the paper.

2. Language Related Tasks and Challenges in the Legal Domain

The way legal language is represented may differ substantially from everyday language [14]. Regulatory texts and to some extent contracts are characterised by complex sentences including numbered lists, citations, references, and special nomenclature. Latin text may also infiltrate legal documents. In legal practice, some authors differentiate between issue centric and document-centric tasks [15]. When drafting documents, the work involves selecting appropriate elements (clauses, paragraphs, etc.) and then filling into the templates the values related to a specific customer. While in issue centric problems an additional so-called presentation layer is required to represent the rules relevant to any related decisions, in which case those selected rules need to be managed as well. In the former case, it is further possible that elements of the documents are selected based on the values provided leading to a procedural approach [16].

Processes cover the (elementary) tasks and the roles fulfilling those tasks in the legal workflow. Restructuring any part of legal workflows may be supported or even initiated by infocommunication technologies. The number of types of workflows and tasks in them may be

higher for more complex or specialized legal organisations. Most core legal workflows contain tasks (steps) that either deal with legal documents, work on written legal statements (such as from letters or emails) or explore the legal domain to solve a problem. For example, reviewing documents for discovery is not a process with simple yes or no answers, and the unique context of the case often determines the degree of relevance for each document. The question of ICT use then relates to which workflow steps may be automated, augmented, or left fully for humans.

Language is central in both law-making and during the application of the law [17, 1, 14, 18]. Indeed, both private and public sector entities are affected – albeit quite differently. Legal reasoning and argumentation are a specific set of skills, and although they are pragmatic, they may ignore the theoretical framework of formal logic. Besides strict logic, argumentation by analogy and examples is an important part of the legal reasoning toolset. Lawyers must argue for the rules themselves and show why a particular rule (or major premise) should apply to a particular case. Law is inherently indeterminate because valid but contradictory legal arguments potentially exist regarding the interpretation of the law, and legal arguments are often arguments about what the language means or ought to mean. There are considerable constraints on what kind of wording is acceptable in legal texts and if the linguistic layer of the template text is not sufficiently abstract, then other (such as business logic related) tools might need to be used for linguistic corrections. This will slow down the creation of the templates and increase the lifetime cost of the software. Although on the positive side, only a small number of typos are expected in legal texts, except perhaps for raw drafts, a rare training set. The morphological diversity of legal language is certainly smaller than everyday speech. It lacks informal addressing and use of first and second persons, and there are fewer verbs. On the other hand, legal texts often incorporate other professions' special terms.

3. NLP and its Importance in Legal Contexts

Natural Language Processing is a special AI technology allowing for innovative text or speech-based solutions. It covers "*a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications*" ([11], p. 2). NLP is typically split into two areas: natural language understanding (NLU) and natural language generation (NLG). The former implies analysing text or speech to infer meaning, while the latter covers using data to generate text or speech that reads or sounds like human (or at least close). Related areas are speech recognition and character recognition which generate input for NLU.

Professional sectors that are perhaps subject to less regulation have been utilizing NLP technologies for some time. The most visible applications involve machine translation solutions (using NLU) and chatbots, where the latter being most prevalent in customer helplines using both NLU and NLG [19]. Other specific applications of text generation also include word-processing applications, transcribing data points into text, or bulk emailing and mail merging functions as well as writing somewhat repetitive news pieces such as weather forecasts based on meteorology data or financial reports generated from stock market data [20]. Less fancy options include question-and-answer systems (which can find relevant information) or tools

that can create summaries of longer texts. The techniques behind NLP have changed a lot over the decades and most recent developments rely on various forms of machine learning including complex, dedicated multi-layer neural networks based structures, such as BERT or GPT-3 (which are designed to consider the context and location of phrases instead of just focusing on words and short phrases). However, these advanced techniques require large amounts of data, they are computationally very intensive (especially during training), therefore, they need not only significant architectural and financial investment but also special expertise to build. If the general language model is to be applied to a specific task, it also needs to be fine-tuned to the specific target texts.

NLP may help in many tasks of the various legal workflows, but there are some challenges as well. Considering the characteristics of the legal language for AI training, sentences and paragraphs are important clues. In case of legal text, this points to a potential de novo pre-training rather than cross-domain training from everyday language. Furthermore, if the complexity of using a tool goes beyond a certain level, that might be too much to ask from legal professionals to deal with. One possibility is to use clause-level NLG instead of full template level as that would help reusability and would fit the logic of legal well. NLP may be used in the creation of contract documents based on templates, which can be very efficient in case of a larger customer base. It can also support legal workflows of document assembly and drafting [21]. Such workflows have two key parts: the creation of the template and its application. However, there is no compatibility between templates of products offered by different vendors – i.e., there is no standardisation in this field – which could cause a problem if users face limitations of a tool after investing a lot of effort to create templates. The creation of templates might require dedicated skills and knowledge of the given tool – which could be expected from a larger organisation but could be challenging for an individual practising lawyer (even with an assistant). Indeed, commercially available tools are often linguistically and legally (i.e., semantically) independent solutions, which makes their use inflexible. It is also possible, considering the diverse language and legal domain options, that different use cases require different products. In that case, such document assembly products might not be attractive for a larger customer base due to the price and learning curve required. Furthermore, the logical structuring of legal messages – especially in their verbal form – may hamper the application of general automated reasoning algorithms in this field. In principle, all NLP tools and methods can be applied in both common law and public law domains. In a common law context, it is vital to find all relevant judicial opinions, whereas in civil law systems codified statutes predominate. As to applying named entity recognition (NER), finding similarities in text corpora such as among various cases and norms, or comparing the dates of the origin of texts, the methods are similar.

Considering the public sector, AI and specifically NLP-based solutions may support law-making and may be applied in citizens' dealings with public entities (government, municipalities, administrative agencies) – in other words, both the legislative and the administrative function of the state may utilize AI and NLP [17, 22, 23]. But one should not forget about one of the most controversial uses of AI, the role of the judiciary, such as judge 'support systems' [24, 25]. One of the most popular applications is to use of chatbots for citizen communication [26]. These digital representatives could be either auditory or textual computerized conversational systems and may be used to provide citizen services through answering questions, routing service requests, handling complaints, supporting form selection or offering translations [27].

The obvious advantage is efficiency, as they can reduce the workload of help centre operators (in cases of administrative questions). NLU can be utilized in searching documents while NLG is applied in drafting documents. Legislative NLP applications include drafting the legislative text, doing a syntactic check of drafts, or summarising long proposals. In the judicial segment, NLG solutions can be used to support judges' decisions: they may help to explain the decision and their results are easier to maintain [15]. The combined use of named NER and relation mining in judicial decisions can help find documents where a person was present in a particular role (e.g., as a defendant) without the need to manually build a database by pre-processing documents beforehand. The issue- or discourse-based approaches cover judge-focused use cases - the set of legal rules represented and described in IT tools determine the questions a user has to answer during the use [28].

4. The Small Language Problem and its Challenge in the Legal Profession

While a typical NLG solution might require dozens of major users and tens of thousands of requests to be generated annually to make it financially feasible, it might be a tough sell for markets of a few million speakers which is normal for many small languages. In other words, the question is related to the size of the population behind a certain language, leading to key differences between big and small languages. In the European Union, for example, only 5 of its 27 members have more than 20 million inhabitants (speaking the same language), and 15 have a population below 10m. This issue is not specific to the EU only but impacts several Asian and African countries as well.

Machine learning tasks always require large text corpora. While a general language model can be built by collecting only the Wikipedia entries for a given language, a more specific task (e.g., legal NER) requires specific corpora. For small languages, there is often no corpus available for more specific tasks, or no publicly available corpus exists from which, for example, a fine-tuning of a model can be trained. Indeed, AI tools such as BERT or GPT-3 require that a comprehensive set of authentic, high-quality legal texts be provided. In addition, these may have to be annotated manually, which may require the involvement of linguists and IT experts. The availability of such expertise may be more limited for languages with fewer speakers. The key problem here is that due to the size of the market it is not immediately economically feasible to develop NLP technologies as they require considerable time and effort to establish a sufficient language-specific base. Even with an acceptable starting corpus, domain or profession-specific frameworks (such as a legal corpus) would also be needed before the final development of a particular application or solution. Legal firms, departments and other legal organisations working with languages of small populations, and therefore small markets often suffer heavy disadvantages compared to languages of large populations. One additional challenge considers linguistic accuracy: although beyond a certain point that is not necessarily a high priority to all applications of NLP, in the legal domain it is crucial.

Solutions available for large languages may not be reachable, and even those face many additional challenges [12]. Although legal advice and content portals gain momentum in several countries, especially in common law context, their business model is not directly sustainable in

smaller jurisdictions. The same holds for cloud services (similarly to general Legal Tech). It is also common for some template authoring functionality to appear partly on a web interface (running on a separate server) and partly on the end-user machine. While these show steady growth, they do not seem to get traction in smaller countries [29].

5. Solutions for the Small Language Problem and the Role of Governments

As certain techniques require dozens of terabytes of training data which is unlikely to be available for the legislative, legal, and judicial practice of smaller languages, various innovative solutions have been tried. When the quantity of training text is insufficient for AI (thus the resulting model would be underfitted), a two-step procedure termed transfer learning is usually recommended. Cross-language Transfer Learning (CLTL) is one of the technologies that can potentially alleviate small language impediments. In cross-language transfer learning machine knowledge is transferred from a (resource-rich) source language to another, target (resource-poor) language. Resources in this regard are annotated text corpora and examples. The same-language and cross-language approaches often come mixed, even the training texts may be of different languages. The line between cross-lingual and cross-domain transfer learning is faint. Some of the solutions developed to remedy language challenges faced by legal NLP solutions developed for polycentric languages may also be applicable for small languages as well. A language is called polycentric if it has several (albeit often interacting) codified standard forms (such as various forms of English or French). A study of a tool called LegalBERT [30] has shown that good accuracy (better results) may be achieved in certain tasks if the pre-training itself is done on legal texts, compared to only fine-tuning over an existing general model using some legal text.

While at the core the small language problem has technical and economic challenges, it is worthwhile to consider some recommendations for policymakers. Governments and policy-makers have great ability to influence technological change in general and it is no different for AI or NLP in particular [22]. Public entities appear both as users and supporters of new technologies. Promoting technologies may be direct or indirect. For the latter, it is possible to support technologies (such as NLP) through infrastructure services or using special award criteria in public procurement spending. More direct influence could be exercised through regulations and financing. Grants may be established to support investors and developers. Depending on the cultural and economic context it is often customary to operate public-funded research in this kind of situation, well market players might not find it feasible to invest in related R&D. Beyond promoting the agenda and open resources, government entities may protect this arena as well by creating alternative options or avoiding monopolistic situations. They may also apply NLP solutions themselves to create demand. Overall, public players should support progress while controlling and overseeing the NLP field at the same time.

Using NLP technologies in the public sector is growing (as evidenced earlier, such as chatbots), and investing through use is probably the best option for a good return. Regulating NLP and the use of NLP in the legal field should be exercised cautiously. The issue is likely to be country dependent as it is related to how big the market is and is further complicated by the size of the

language population. Thus, as a first step, it would be important to assess individual situations. It is recommended that a detailed survey discovers the status of language models generally, and legal ones specifically. In the EU, programs could be established for more advanced countries to support others in the development of software and language models. Even more important is to establish technology transfer competence centres regionally to train, consult and support SLF staff. Legal Tech education should be introduced, or the level thereof substantially enhanced in all EU universities. Collaboration among universities should be supported in all known ways. Another important aspect is the availability of free software. Indeed, governments and public agencies often have the responsibility to provide support where market forces may not be able to provide acceptable level solutions. Thus, they may embrace an open-source approach and dedicate resources to developing pools of NLP software [29] that may be freely reused in developing applications for smaller, non-central languages.

In many legal cases, unequal parties face each other. This is especially the situation when a legal office faces a public attorney or a tax authority. The latter have all the documents of all their cases and procedures at hand at a national level; the former may have access to a limited set. This creates an inequality of possibilities to train AI, among others. Therefore, governments should be urged to publish all legal documents reasonably publishable, including court decisions, initiatives and interim documents as well. The level of availability of such documents varies widely, with almost 4m court decisions published in Slovakia [31], and 0.17m in Hungary [32] for example.

Even with the publication of court proceedings, cases resolved outside the court (e.g. by plea bargain [33]) may not reach the public at all. Most contracts will also remain obscure by nature. Governments and professional associations may want to organize, regulate and facilitate the anonymized publication of these documents. Still, if only a single entity sells electricity in a given country, it may not be easy to mask their identity, or, more importantly, the identity of the other party of the contract, without hiding the subject (in short, scarcity is the enemy of anonymisation [34]).Also, governments should support AI training of specific document sets, and the development of applications based on these AI entities, with special care to mask non-public identities from result sets.

6. Summary

This paper has provided a comprehensive review of the so-called small language problem (sometimes approached as the challenge of non-central languages). The key message is the small language problem has a special connotation in the legal field as this profession relies on language heavily – in all forms of communication. One of the key challenges countries and jurisdictions of small languages face in light of increasing NLP applications is the lack of advanced software solutions and the required (foundational) language corpus. While research in the area exists and solutions from larger markets may be applied to a certain extent, the local market is often not strong enough to create high quality, specialized solutions. Consequently, the main message of this report is to call for support from the public sector. Government (or other public entities) should step forward and carry the flag both in financial support and by being a user themselves.

References

[1] R. Dale, Law and word order: Nlp in legal tech, Natural Language Engineering 25 (2019) 211–217.

[2] M.-M. Bues, E. Matthaei, Legaltech on the rise: Technology changes legal work behaviours, but does not replace its profession, in: Liquid Legal, Springer, 2017, p. 89–109.

[3] K. Mania, Legal technology: Assessment of the legal tech industry's potential, Journal of the Knowledge Economy (2022) 1–25.

[4] J. Frankenreiter, J. Nyarko, Natural language processing in legal tech, Legal Tech and the Future of Civil Justice (David Engstrom Ed.). (2022). URL: https://dx.doi.org/10.2139/ssrn.4027030. doi:10.2139/ssrn.4027030.

[5] R. Sil, A. Roy, B. Bhushan, A. Mazumdar, Artificial intelligence and machine learning based legal application: The state-of-the-art and future research trends, in: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS, 2019, p. 57–62.

[6] M. Katsh, Competing in cyberspace: The future of the legal profession, Technological Forecasting and Social Change 52 (1996) 109–117.

[7] M. Corrales, M. Fenwick, H. Haapio, Legal Tech, Smart Contracts and Blockchain, Springer, 2019.

[8] D. Lewis, Afterword: Data, knowledge, and e-discovery, Artificial Intelligence and Law 18 (2010) 481–486.

[9] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How does nlp benefit legal system: A summary of legal artificial intelligence, 2020.

[10] M. McKamey, Legal technology: Artificial intelligence and the future of law practice, Appeal: Rev. Current L. & L. Reform 22 (2017) 45.

[11] E. Liddy, Natural language processing, 2001.

[12] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6282–6293. URL: https://aclanthology.org/2020.acl-main.560. doi:10.18653/v1/2020.acl-main.560.

[13] Y. Zhang, A. Warstadt, H.-S. Li, S. Bowman, When do you need billions of words of pretraining data?, 2020.

[14] M. Legg, F. Bell, Artificial intelligence and the legal profession: Becoming the ai-enhanced lawyer, U. Tas. L. Rev 38 (2019) 34.

[15] L. Branting, An issue-oriented approach to judicial document assembly, in: Proceedings of the 4th International Conference on Artificial Intelligence and Law, 1993, p. 228–235.

[16] L. Branting, C. Callaway, B. Mott, J. Lester, Integrating discourse and domain knowledge for document drafting, in: Proceedings of the 7th International Conference on Artificial Intelligence and Law, 1999, p. 214–220.

[17] G. Buchholtz, Artificial intelligence and legal tech: Challenges to the rule of law, in: Regulating artificial intelligence, Springer, 2020, p. 175–198.

[18] T. Wischmeyer, T. Rademacher (Eds.), Regulating Artificial Intelligence, Springe Nature, 2020.

[19] D. Jurafsky, J. Martin, Speech and language processing (draft), 2018. URL: https://web.stanford.edu/~jurafsky/slp3/.
[20] N. Indurkhya, F. Damerau, Handbook of natural language processing, Chapman and Hall/CRC, 2010.
[21] M. Lauritsen, Current frontiers in legal drafting systems, in: Proceedings of the 11th International Conference on AI and Law, 2007.
[22] C. Djeffal, Artificial intelligence and public governance: Normative guidelines for artificial intelligence in government and public administration, in: Regulating Artificial Intelligence, Springer, 2020, p. 277–293.
[23] D. Hogan-Doran, Computer says" no": Automation, algorithms and artificial intelligence in government decision-making, Judicial Review: Selected Conference Papers: Journal of the Judicial Commission of New South Wales, The 13 (2017) 345–382.
[24] O. Abiodun, A. Lekan, Exploring the potentials of artificial intelligence in the judiciary, International Journal of Engineering Applied Sciences and Technology 5 (2020) 23–27.
[25] K. Forrest, When Machines Can Be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence, World Scientific, 2021.
[26] E. Tambouris, Using chatbots and semantics to exploit public sector information, EGOV-CeDEM-EPart (2018) 125.
[27] H. Mehr, H. Ash, D. Fellow, Artificial intelligence for citizen services and government, Ash Cent. Democr. Gov. Innov. Harvard Kennedy Sch (2017) 1–12.
[28] M. Marković, S. Gostojić, Knowledge-based legal document assembly, 2020. URL: https://arxiv.org/abs/2009.06611.
[29] O. Streiter, K. Scannell, M. Stuflesser, Implementing nlp projects for noncentral languages: Instructions for funding bodies, strategies for developers, Machine Translation 20 (2006) 267–289.
[30] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, 2020.
[31] D. Varga, Z. Szoplák, S. Krajci, P. Sokol, P. Gurský, Analysis and prediction of legal judgements in the slovak criminal proceedings, Information Technologies – Applications and Theory 2962 (2021) 161–170.
[32] G. Görög, P. Weisz, Legal entity recognition in an agglutinating language and document connection network for eu legislation and eu/hungarian case law, 2019. arXiv:1907.12280.
[33] A. Gácsi, The plea bargaining in hungary. scientific works of national aviation university, Series: Law Journal 'Air and Space Law' 3 (2018) 166–176. URL: https://doi.org/10.18372/2307-9061.48.13203. doi:10.18372/2307-9061.48.13203.
[34] G. Csányi, D. Nagy, R. Vági, J. Vadász, T. Orosz, Challenges and open problems of legal document anonymization, Symmetry 13 (2021).