Proceedings Paper:
Maynard, D. orcid.org/0000-0002-1773-7020 and Funk, J. (2020) Combining expert knowledge with NLP for specialised applications. In: Sojka, P., Kopeček, I., Pala, K. and Horák, A., (eds.) Text, Speech, and Dialogue: 23rd International Conference on Text, Speech and Dialogue (TSD 2020). TSD 2020 - 23rd International Conference on Text, Speech and Dialogue, 08-11 Sep 2020, Brno, Czech Republic. Lecture Notes in Computer Science, 12284 . Springer , pp. 3-10. ISBN 9783030583224

https://doi.org/10.1007/978-3-030-58323-1_1

Combining expert knowledge with NLP for specialised applications*

Diana Maynard[0000−0002−1773−7020] and Adam Funk[0000−0001−8404−1173]

Dept. of Computer Science, University of Sheffield, Sheffield, UK
d.maynard@sheffield.ac.uk

Abstract. Traditionally, there has been a disconnect between custom-built applications used to solve real-world information extraction problems in industry, and automated learning-based approaches developed in academia. Despite approaches such as transfer-based learning, adapting these to more customised solutions where the task and data may be different, and where training data may be largely unavailable, is still hugely problematic, with the result that many systems still need to be custom-built using expert hand-crafted knowledge, and do not scale. In the legal domain, a traditional slow adopter of technology, black box machine learning-based systems are too untrustworthy to be widely used. In industrial settings, the fine-grained highly specialised knowledge of human experts is still critical, and it is not obvious how to integrate this into automated classification systems. In this paper, we examine two case studies from recent work combining this expert human knowledge with automated NLP technologies.

Keywords: Natural Language Processing · ontologies · information extraction

1 Introduction

Although machine learning, and more recently deep learning-based approaches, have shown enormous promise and success in Natural Language Processing (NLP), and more generally in the field of Artificial Intelligence (AI), there are nevertheless a number of drawbacks when applied to many real-world applications in industrial settings. The medical and legal domains have been traditionally slow to adopt automated technologies, due partly to the critical effect of mistakes. On the other hand, driverless cars and autonomous robots are fast becoming an everyday reality, despite the numerous ethical considerations. When a human driver hits the brakes in order to avoid hitting a child who runs in front of a car, they make a moral decision to shift the risk from the child to their passengers. How should an autonomous car react in such a situation? One piece of research [3] showed that in surveys, people preferred an autonomous vehicle

* This work was partially supported by European Union under grant agreement No.726992 KNOWMAK and No. 825091 RISIS; and by Innovate UK.

to protect pedestrians even if it meant sacrificing its passengers, as most human drivers would do, but paradoxically, these people claimed that they would not want to buy one if it were programmed to do so.

The recent COVID-19 pandemic has driven a wealth of interest in automated AI technology such as call systems. While call centers have long been a forerunner in the use of such tools, the pandemic has accelerated their growth due to the combination of a shortage of workers and an enormous increase in calls. IBM witnessed a 40% increase in use of Watson Assistant between February and April 2020, and other technologies show a similar popularity rise.[1]

However, automated call systems only deal with part of the problem, and are still relatively simple. They are best at signposting users to sources of information and mostly rely on posing pre-set questions with simple answers that can be easily be processed (e.g. yes/no questions, or by spotting simple keywords). Adapting these kinds of conversational agents to the specific demands of individual businesses requires intensive labour and training materials, so is not a project to be undertaken lightly or urgently.

In this paper, we focus on two case studies in which we have investigated how expert human knowledge can be interlinked with the advantages of automated technologies. These enable traditional manual tasks to be carried out faster and more accurately by processing huge amounts of data, while still ensuring both the consistency and flexibility to deal with new data as it emerges. The first of these is in the legal domain, where we have developed tools to assist consultants to review collateral warranties - an expensive and time-consuming task which nevertheless demands high precision and intricate levels of linguistic detail. The second is in the wider field of European scientific and technological knowledge production and policy making, where tools are needed to assist policymakers in understanding the nature of this enormous, highly complex and fast-changing domain.

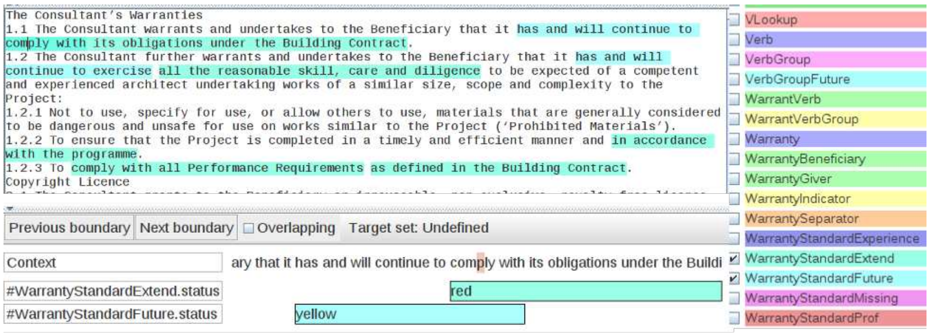
2 Legal IE

The reviewing of collateral warranties is an important legal and economic task in the construction industry. These warranties are a type of contract by which a member of the construction team (e.g. an architect) promises a third party (e.g. the project funder) that they have properly discharged their contract. For example, an architect of a new office development owes a duty of care to the occupier of the development, concerning any design defects that might show up later. Without a collateral warranty, the architect would typically not be liable. Collateral warranties may include ‘step-in’ rights which allow the beneficiary to step into the role of the main contractor. This can be important, for example to banks providing funding for a project, enabling them to ensure that the project is completed if that contractor becomes insolvent.

[1] https://www.technologyreview.com/2020/05/14/1001716/ai-chatbots-take-call-center-jobs-during-coronavirus-pandemic

There are a number of standard forms of collateral warranty, but their specific terms can be disputed, with clients often claiming that industry standard warranties favour subcontractors and designers. There may also be complex wording or terminology in standard contracts which make them too risky because they are outside the scope of the warranty giver's insurance cover. Therefore, many collateral warranties are bespoke. However, completing collateral warranties to the satisfaction of all parties is incredibly difficult, especially for large projects with many consultants and sub-contractors, as well as multiple occupants, and it is legally complex and onerous for lawyers to review them. A single manual review typically takes 3 hours, but is often not properly valued by clients, who see it as a sideline to the main construction contract.

We have therefore been developing prototype software to assist lawyers in reviewing collateral warranties. The legal industry typically does not make use of automated software for these kind of tasks. Existing contract review software is limited and based on machine learning, which tends to be inadequate because it neither analyses collateral warranties to the level of detail required, nor does it provide explanatory output. Furthermore, it is unclear how the highly specialised human expertise can be replicated in an automated approach. For this reason, our system uses a rule-based approach which automates some of the more straightforward parts of the review process and focuses on breaking the documents down into relevant sections pertaining to each kind of problem the human reviewer must address. It uses a traffic light system to check standard protocols and to flag possible problems that the lawyer should investigate, with explanations as to the nature of the problem.



The Consultant's Warranties
1.1 The Consultant warrants and undertakes to the Beneficiary that it has and will continue to comply with its obligations under the Building Contract.
1.2 The Consultant further warrants and undertakes to the Beneficiary that it has and will continue to exercise all the reasonable skill, care and diligence to be expected of a competent and experienced architect undertaking works of a similar size, scope and complexity to the Project:
1.2.1 Not to use, specify for use, or allow others to use, materials that are generally considered to be dangerous and unsafe for use on works similar to the Project ('Prohibited Materials').
1.2.2 To ensure that the Project is completed in a timely and efficient manner and in accordance with the programme.
1.2.3 To comply with all Performance Requirements as defined in the Building Contract.
Copyright Licence

Context: ary that it has and will continue to comply with its obligations under the Buildi
#WarrantyStandardExtend.status: red
#WarrantyStandardFuture.status: yellow

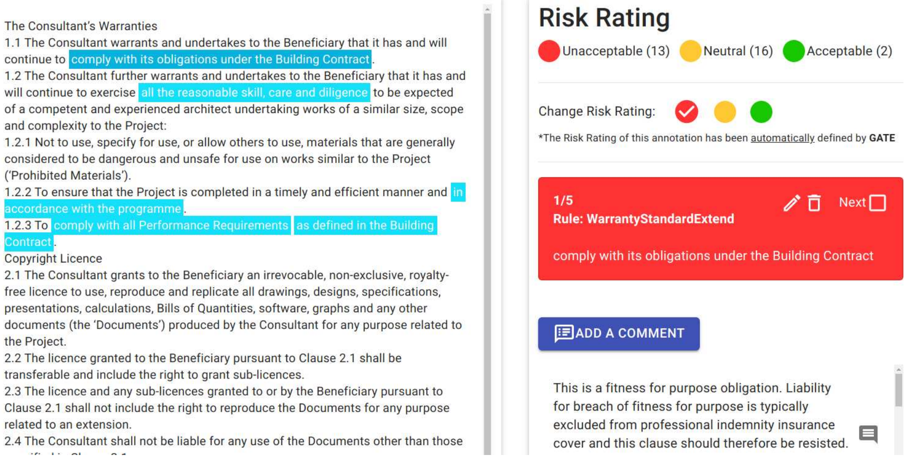
- ☐ VLookup
- ☐ Verb
- ☐ VerbGroup
- ☐ VerbGroupFuture
- ☐ WarrantVerb
- ☐ WarrantVerbGroup
- ☐ Warranty
- ☐ WarrantyBeneficiary
- ☐ WarrantyGiver
- ☐ WarrantyIndicator
- ☐ WarrantySeparator
- ☐ WarrantyStandardExperience
- ☑ WarrantyStandardExtend
- ☑ WarrantyStandardFuture
- ☐ WarrantyStandardMissing
- ☐ WarrantyStandardProf

Fig. 1. Sample annotations in a collateral warranty in the GATE GUI

The warranty annotation tool is based on the GATE architecture for NLP [4], an open source toolkit which has been in development at the University of Sheffield for more than 20 years. A rule-based approach is used to annotate different sections of the document and to recognise certain relevant entities (such as copyright issues, the warranty beneficiary, the warranty giver, and so on). Figure 1 shows an example of a mocked-up warranty annotated in the GATE

GUI.[2] Two annotations are highlighted here, which concern the extent of the warranty standard and the future warranty standard. In the bottom part of the picture, we see that these have features *red* and *yellow* respectively. This indicates that this part of the contract is something that a human reviewer needs to check manually.

The human reviewer does not see the GATE GUI at all; we show it only to explain the underlying technology. Instead, they use the reviewing interface also developed in the project, which enables them to upload a document, select some parameters, and run GATE on it via a web service. They can then view the contract in the interface and zoom in on different parts of the document to see the suggestions and highlights that GATE has made in an easily understandable way. The yellow and red flags ("translated" from the GATE features) indicate that they need to review these parts, and the review cannot be marked as completed until these are satisfactory. Figure 2 shows the same mocked-up document now in the reviewing interface. The reviewing process semi-automatically generates a final report (for the lawyer's client) based on the current human-written report, with warnings about the risky passages in the document.



The Consultant's Warranties
1.1 The Consultant warrants and undertakes to the Beneficiary that it has and will continue to comply with its obligations under the Building Contract.
1.2 The Consultant further warrants and undertakes to the Beneficiary that it has and will continue to exercise all the reasonable skill, care and diligence to be expected of a competent and experienced architect undertaking works of a similar size, scope and complexity to the Project:
1.2.1 Not to use, specify for use, or allow others to use, materials that are generally considered to be dangerous and unsafe for use on works similar to the Project ('Prohibited Materials').
1.2.2 To ensure that the Project is completed in a timely and efficient manner and in accordance with the programme.
1.2.3 To comply with all Performance Requirements as defined in the Building Contract.
Copyright Licence
2.1 The Consultant grants to the Beneficiary an irrevocable, non-exclusive, royalty-free licence to use, reproduce and replicate all drawings, designs, specifications, presentations, calculations, Bills of Quantities, software, graphs and any other documents (the 'Documents') produced by the Consultant for any purpose related to the Project.
2.2 The licence granted to the Beneficiary pursuant to Clause 2.1 shall be transferable and include the right to grant sub-licences.
2.3 The licence and any sub-licences granted to or by the Beneficiary pursuant to Clause 2.1 shall not include the right to reproduce the Documents for any purpose related to an extension.
2.4 The Consultant shall not be liable for any use of the Documents other than those

Risk Rating

● Unacceptable (13) ● Neutral (16) ● Acceptable (2)

Change Risk Rating: ✓ ● ●

*The Risk Rating of this annotation has been automatically defined by **GATE**

1/5
Rule: WarrantyStandardExtend

comply with its obligations under the Building Contract

This is a fitness for purpose obligation. Liability for breach of fitness for purpose is typically excluded from professional indemnity insurance cover and this clause should therefore be resisted.

Fig. 2. Sample annotations in a mock-up collateral warranty in the GATE GUI

3 Understanding scientific knowledge production in Europe

Understanding knowledge production and co-creation in key emerging areas of European research is critical for policy makers wishing to analyse impact and

[2] The warranty is not a real one, for legal reasons, but the annotation is genuine.

make strategic decisions. Essentially, they need to know who is doing research on what topic and in which country or region. The RISIS-KNOWMAK tool[3] is the result of a 3-year European project enabling the user to combine multiple data sources (publications, patents, and European projects), connect the dots by analysing knowledge production by topics and geography, and to pick from different kinds of visualisation options for the data they are interested in.

The tool generates aggregated indicators to characterise geographical spaces (countries or regions) and actors (public research organisations and companies) in terms of various dimensions of knowledge production. For each topic or combination of topics, the mapping of documents enables the generation of indicators such as the number of publications, EU-FP projects, and patents in a specific region, as well as various composite indicators combining dimensions, such as the aggregated knowledge production share and intensity, and the publication degree centrality.

Current methods for characterising and visualising the field have limitations concerning the changing nature of research, differences in language and topic structure between policies and scientific topics, and coverage of a broad range of scientific and political issues that have different characteristics. The kind of language used in patent descriptions is very different from that used in scientific publications, and even the terminology can be very different, so it is hard to develop tools which can classify both kinds of document in the same way.

In recent years, a priori classification systems for science and technology, such as the Field of Science Classification (OECD, 2002) and IPC codes for patents [6], have been increasingly replaced by data-driven approaches, relying on the automated treatment of large corpora, such as word co-occurrences in academic papers [2], clustering through co-citation analysis [9], and overlay maps to visualise knowledge domains [7]. These approaches have obvious advantages, since they are more flexible to accommodate the changing structures of science, and are able to discover latent structures of science rather than impose a predefined structure over the data [8]. Yet, when the goal is to produce indicators for policymakers, purely data-driven methods also display limitations. On the one hand, such methods provide very detailed views of specific knowledge domains, but are less suited to large-scale mapping across the whole science and technology landscape. On the other hand, lacking a common ontology of scientific and technological domains [5], such mappings are largely incommensurable across dimensions of knowledge production. Perhaps even more importantly, data-driven methods do not allow presumptions of categories used in the policy debate to be integrated in the classification process. These are largely implicit and subjective, implying that there is no gold standard against which to assess the quality and relevance of the indicators, but these are inherently debatable [1].

The RISIS-KNOWMAK classification tool is a GATE-based web service which classifies each document according to the relevant topics it is concerned with. This involves the novel use of ontologies and semantic technologies as a means to bridge the linguistic and conceptual gap between policy questions

[3] https://www.knowmak.eu/

and (disparate) data sources. Our experience suggests that a proper interlinking between intellectual tasks and the use of advanced techniques for language processing is key for the success of this endeavour.

Our approach was based on two main elements: a) the design of an ontology of the Key Enabling Technologies and Societal Grand Challenges (KET and SGC) knowledge domains to make explicit their content and to provide a common structure across dimensions of knowledge production; and b) the integration between NLP techniques (to associate data sources with the ontology categories) and expert-based judgement (to make sensible choices for the matching process). This drove a recursive process where the ontology development and data annotation were successively refined based on expert assessment of the generated indicators.

Ontology development in our application involves three aspects: first, the design of the ontology structure, consisting of a set of related topics and subtopics in the relevant subject areas; second, populating the ontology with keywords; and third, classifying documents based on the weighted frequency of keywords. The mapping process can be seen as a problem of multi-class classification, with a large number of classes, and is achieved by relying on source-specific vocabularies and mapping techniques that also exploit (expert) knowledge about the structure of individual data sources. This is an iterative process, based on co-dependencies between data, topics, and the representation system.

Our initial ontology derived from policy documents was manually enriched and customised, based on the outcome of the matching process and expert assessment of the results. Eventually, the original ontology classes may also be adapted based on their distinctiveness in terms of data items. Such a staged approach, distinguishing between core elements that are stabilised (the ontology classes) and elements that are dynamic and can be revised (the assignment of data items to classes), is desirable from a design and user perspective. Therefore, the approach is flexible, for example to respond to changes in policy interests (see Section 5), and scalable since new data sources can be integrated within the process whenever required. All three steps require human intervention to define prior assumptions and to evaluate outcomes, but they integrate automatic processing through advanced NLP techniques. Consequently, if changes are deemed necessary, the process can easily be re-run and the data re-annotated within a reasonable period of time.

The ontology is freely available on the project web page[4]; we refer the interested reader also to the publications and documentation found there for full details of the technology. Our experience with this specialised ontology and classification shows that while NLP techniques are critical for linking ontologies with large datasets, some key design choices on the ontology and its application to data are of an intellectual nature and closely associated with specific user needs. This suggests that the design of interactions between expert-based a priori knowledge and the use of advanced data techniques is a key requirement for robust S&T ontologies.

[4] https://gate.ac.uk/projects/knowmak/

We have also produced a number of case studies of how the tool could be used for policy making. In the field of genomics, we compared the technological and scientific knowledge production in Europe in the period 2010-2014. Technological production is measured by patents, while scientific production is measured by publications. These show different geographical distributions. The former is more concentrated in space: in terms of volume, Paris is the biggest cluster for both types. Within regions, production varies a lot: London is the biggest producer of both types, while Eindhoven is key in terms of technological knowledge (both for volume and intensity). These findings clearly reflect the different structure of public and private knowledge.

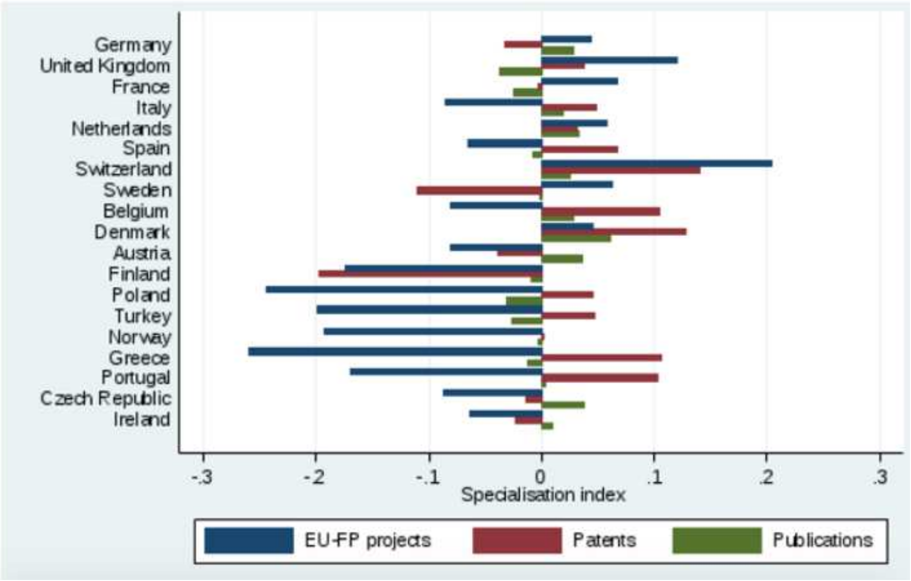


Fig. 3. Specialisation indexes in biotechnology around Europe

Another example is based on the topic of Industrial Biotechnology (IB), which offers new tools, products and technologies based on the exploitation of biological processes, organisms, cells or cellular components. Policymakers might like to know, for example, which European countries are (more) specialised in this field, and whether there are differences in the extent of specialisation when considering scientific and technological development. The tool provides ready-to-use indicators to answer these questions. Figure 3 indicates the country specialisation indexes in biotechnology for the three measures of knowledge production in the period 2010-2014. Values greater/lower than 0 in the specialisation indexes imply that a country is more/less specialised in IB compared with the average European country. Amongst larger countries in terms of knowledge production, Germany, France, Italy and the Netherlands exhibit no clear specialisation in IB, with all indexes ranging at moderate levels from -0.09 to 0.07. The only exception is the UK, which is more specialised in terms of EU-FP projects (specialisation higher than 0.1).

4 Conclusions

This paper has focused on two case studies based around tools we have developed for specialised applications (in the legal and scientometrics domains) where standard NLP tools based on machine learning are unlikely to be satisfactory due to the kinds of knowledge and output required, and to other constraints such as explainability (in the legal case) and flexibility (in the scientometrics case). While new advances in deep learning continue to transform the levels of achievement of automated tools for a number of NLP classification tasks, as well as in machine translation and in speech and image recognition, nevertheless they are not suitable for all NLP tasks, at least as stand-alone tools. Rule-based systems and the incorporation of human expert knowledge interweaved with advanced learning may provide better approaches in some cases, as we have demonstrated. Important future directions in the field of NLP lie not only in improving the explainability of machine learning tools, such as with the use of adversarial examples, and improved linguistic knowledge in neural networks, but also in investigating more deeply the ways in which expert knowledge can best be integrated.

References

1. Barré, R.: Sense and nonsense of s&t productivity indicators. Science and Public Policy **28**(4), 259–266 (2001)
2. Van den Besselaar, P., Heimeriks, G.: Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. Scientometrics **68**(3), 377–393 (2006)
3. Bonnefon, J.F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**(6293), 1573–1576 (2016)
4. Cunningham, H.: Gate, a general architecture for text engineering. Computers and the Humanities **36**(2), 223–254 (2002)
5. Daraio, C., Lenzerini, M., Leporelli, C., Moed, H.F., Naggar, P., Bonaccorsi, A., Bartolucci, A.: Data integration for research and innovation policy: an ontology-based data management approach. Scientometrics **106**(2), 857–871 (2016)
6. Debackere, K., Luwel, M.: Patent data for monitoring s&t portfolios. In: Handbook of quantitative science and technology research, pp. 569–585. Springer (2004)
7. Rafols, I., Porter, A.L., Leydesdorff, L.: Science overlay maps: A new tool for research policy and library management. Journal of the American Society for information Science and Technology **61**(9), 1871–1887 (2010)
8. Shiffrin, R.M., Börner, K.: Mapping knowledge domains (2004)
9. Šubelj, L., van Eck, N.J., Waltman, L.: Clustering scientific publications based on citation relations: A systematic comparison of different methods. PloS one **11**(4) (2016)