Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-141

Using Agreement Statements to Identify Majority Opinion in UKHL Case Law

Josef VALVODA [a], Oliver RAY [a] and Ken SATOH [b]
[a] *University of Bristol, Bristol, UK,*
e-mail: {jv16618,csxor}@bristol.ac.uk
[b] *National Institute of Informatics, Tokyo, Japan,*
e-mail: ksatoh@nii.ac.jp

Abstract. This paper is concerned with the task of finding majority opinion (MO) in UK House of Lords (UKHL) case law by analysing agreement statements (AS) that explicitly express the appointed judges' acceptance of each other's reasoning. We introduce a corpus of 300 UKHL cases in which the relevant AS and MO have been annotated by three legal experts; and we introduce an AI system that automatically identifies this AS and MO with a performance comparable to humans.

Keywords. Agreement Statements, Majority Opinion, UK House of Lords

1. Introduction

The court of the *UK House of Lords* (UKHL) is the former judicial arm of the British parliament's upper house, which served as the country's highest appellate court until it became the *UK Supreme Court* (UKSC) in 2009. In this court, a *majority decision* (MD) is an outcome agreed by more than half of the participating judges, while a *majority opinion* (MO) is a line of reasoning accepted by more than half as legal grounds for that decision. The distinction is crucial because an MO sets a *binding* precedent in UK law, while a non-majority view is merely *persuasive* even if it supports an MD. Thus it is common for UK law lords to discuss their opinions in draft and explicitly state any agreements with each other in their judgments. Usually this is done through formulaic phrases, that we call *agreement statements* (AS), used specifically for this purpose.

The goal of our work is to develop a computational method for detecting AS in UKHL judgments and using them to identify cases with a binding MO. This is needed because legal research tools[1] currently offer little help in this respect: since, unlike other jurisdictions, the obvious instances of dissent which are often flagged up in case digests are generally insufficient to establish the presence or absence of MO in UKHL cases. This paper takes the natural first step towards a solution by looking for unqualified (in-full) AS that suffice (per-se) to imply a definite (non-contestable) MO.

Our first contribution is to introduce a new legal corpus, called ASMO, consisting of 300 UKHL cases in which relevant AS and implied MO have been annotated by

[1] See for example westlaw.co.uk , lexisnexis.com/uk/legal , justcite.com and bailii.org/uk/cases/ukhl/

three legal experts. We derive a consensus labeling for the corpus and determine how accurately humans perform this task in practice. Our second contribution is to introduce a novel AI system that uses *machine learning* (ML) and *natural language processing* (NLP) to identify relevant AS and MO with a performance comparable to humans. We argue this is a useful first step towards the development of practical tools to help lawyers identify legal precedents in UKHL judgments, but we also demonstrate why this task is far more complex than it might first seem.

2. Background

Legal scholars have long discussed [1,2] how the UKHL tradition of publishing *seriatim* opinions of individual judges, with no accompanying statement of official consensus on the underlying reasoning, can make it hard to distinguish a binding MO from a persuasive MD, even when judges use explicit AS to express their agreements with each other.

In a speech [3] on the first anniversary of the UKSC Lady Hale stated while *"there should never be any doubt about what has been decided and why [...] This may not always be achieved even when we think that we have"*. Citing several UKHL & UKSC cases, she singled out the ongoing failure to solve this as the *"low point"* of the year.

Our aim is to approach this problem from a computational perspective by automating the task of detecting unqualified AS in UKHL judgments and using them to identify cases where they are sufficient to establish a definite MO. This can be broken down into two key tasks which are explained in the following two sub-sections.

2.1. Agreement Statements (AS)

The first challenge is to identify those sentences in which the judges actually specify their agreements with each other. Although a stock of formulaic phrases have evolved for this purpose, subtle variations in the precise English wording allow judges to express a myriad range of full or partial agreements - as illustrated in Table 1 below:

1	I fully agree with my noble and learned friend Lord Woolf that this appeal should be dismissed for the reasons he gives.
2	I agree with it, and for the reasons which he gives I, too, would dismiss the appeal.
3	I have read the speeches of my noble and learned friends, Lord Hoffmann and Lord Slynn, and for the reasons given by Lord Hoffmann, I would dismiss this appeal.
4	Therefore, like Lord Hoffmann, I see no reason in principle why, today, prerogative legislation, too, should not be subject to judicial review on ordinary principles of legality, rationality and procedural impropriety.
5	I too would allow the appeal and make the orders my noble friend, Lord Millett, proposes.
6	I have had the advantage of reading in draft the speech of my noble and learned friend, Lord Hoffmann.
7	For these reasons I would allow the appeal and direct that the case be remitted to the County Court for trial.
8	For these reasons and also for those contained in the opinions of my noble and learned friends Lord Scott of Foscote and Lord Brown, I agree with the conclusion reached by Langley J and the Court of Appeal and would dismiss the appeal.
9	I am in full agreement with the reasons expressed in the House today by my noble and learned friends.

Table 1. *Example AS, taken from our corpus, representing Full Agreement (1,2,3,8), Partial Agreement (4), Order Agreement (5), Acknowledgement (3,6), Self Agreement (7,8) and Generic Agreement (9).*

As our aim is to find incontestable MO based on *full* agreements (where one judge accepts another's reasoning in its entirety) our first priority is to distinguish these from weaker *partial* agreements (with just some aspects of the other's reasoning) or *order* agreements (with only the outcomes or orders proposed by the other). For example, in Table 1, sentence 1 is a full agreement with the reasons for the outcome, but sentence 4 is a partial agreement with just a part of the reasoning (relating to prerogative legislation), and sentence 5 is an order agreement with only the outcome and orders (but not the reasoning).

When considering such agreements a few pitfalls must be avoided. For example, sentence 2 is a full agreement where the name of the judge being agreed with is not explicitly contained in the AS, but must be inferred from the relevant pronoun. This is why we must also consider *acknowledgment* statements, like sentence 6, which contain the actual names of the judges referred to. Sentence 3 is a full agreement with one judge combined with the acknowledgement of another. This shows one sentence may contain several AS, and not all the judges mentioned are necessarily being agreed with.

It's worth pointing out the notion of full *agreement* can be more precisely viewed as an *acceptance* that the opinions of some set of judges comprise the binding reasoning of a case. If a judge believes their own reasoning is indispensable, this can be seen as *self* agreement. Although self agreement is often left implicit, judges do frequently refer to their own reasons explicitly, as shown in sentence 7 - especially when they see them as a necessary addition to the reasoning of some other judges, as in sentence 8 (which also includes a partial agreement with a judge in a lower court for good measure).

Finally, it is not uncommon for judges express a full agreement with their *learned friends*, but without saying exactly who. As shown in sentence 9, we call these *generic* agreements. While this judge is likely agreeing with at least two of his peers, we cannot be certain which ones. Could it be all the (other) judges, or just those which have posited arguments of their own, or only those whose arguments are acknowledged by this judge, or something else? But, although generic agreements are ambiguous, if the other judges are more explicit, it may still be possible to find an MO, as explained below.

2.2. Majority Opinions (MO)

The second challenge is to determine if an MO is implied by the AS. For us, this means finding a *set* of judges whose reasoning is collectively agreed with by more than half the court. To this end, we find it convenient to depict the judges of a case as nodes in a graph, using arrows to show full agreements (with loops denoting self agreement and ellipses representing generic agreements) and bold circles to show any judges forming an MO. As most cases in our corpus have 5 judges, we depict them by the letters A-E. In this way, we can use the four hypothetical cases shown in Figure 1 below to illustrate some of the key challenges involved in identifying MO from the AS. In so doing, we will explain the sort of inferences that a lawyer would make - some of which are necessarily based on a familiarity with the way that judges actually express themselves in such cases.

In example (i), a 3-of-5 judge majority {B,C,E} establishes {D} as the MO - since the former all fully agree with the latter's reasoning. But, while the MO is clear, it is worth noting we would also be justified in including D in this majority because the fact he expresses no full agreement with any other judges implies he must be relying on his own arguments. Indeed, it happens in practice that one lord writes a stand-alone *lead*

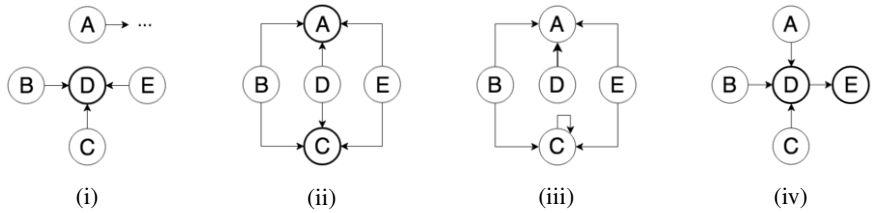


Figure 1. *Graphs showing three hypothetical cases; where letters denote judges, arrows denote full agreement, ellipses denote generic agreement, and bold circles show the judges, if any, whose reasoning forms a MO.*

judgment which all the other judges agree with. Of course, in this example, since A gives only a generic agreement, we cannot actually be sure if he shares the majority view (that the ratio of the case is contained fully in D's opinion) or if he instead believes some additional orbiter from B, C, E, or even A himself, are also indispensable. Fortunately, the MO is not affected by A's position, as D is already supported by a majority here.

In example (ii), a 3-of-5 judge majority {B,D,E} establishes {A,C} as the MO - since all the former agree with the reasoning of both the latter. While the MO is again clear, it is worth noting that neither A nor C are part of the majority electing them as the MO. For, while each certainly agrees with his own reasoning (as neither provides any other external grounds for their decision), we can't assume they agree with each other. It does happen in practice that two judges explicitly express their mutual agreement with each other, but when that occurs they will explicitly agree with each other. Even though that doesn't happen here, a majority nevertheless does still believe that the arguments in the opinions of A and B are both needed to adequately justify the decision.

Example (iii) is nearly identical to the previous one, but shows a case where no MO can be inferred from the AS. The problem is that B and E now form a minority (voting for both A and C), while C forms another minority (voting for C but not A), which means there is no majority view. Note that when a judge agrees with the reasoning of two or more judges, the agreement is with the combined reasoning as opposed to reasoning of one judge or the other. Therefore, the agreement with reasoning of judges A and C, is different from agreement with the reasoning of judge A or C. In our example (iii), B and E are not agreeing with A or C, they agree with both A and C. Crucially, this also shows the MO cannot be determined by simply finding nodes with three or more incoming edges. Example (iv) also demonstrates this point. Here a majority {A,B,C} all agree with D, who in-turn agrees with E. But, since the reasoning of D relies on E, it follows they both must be included in the MO which is therefore {D,E}. Note that, as we only consider full agreements here, A, B and C must implicitly agree with E - since if they did not then they would not be able to agree with D who clearly does. This is another illustration of how hard this task can be.

3. Manual Annotation Study

To better understand the practical significance of the challenges outlined in the previous section, we created a corpus of 300 UKHL judgments and asked three experts to identify the relevant AS and MO in each case. We then used an arbitration process to derive a

consensus annotation for the entire corpus. The following subsections describe how our corpus, called ASMO[2], was constructed, annotated and arbitrated.

3.1. Creating the Corpus

At the outset, we decided a corpus of 300 cases should provide an adequate basis for reliably training and testing an AI system. But, sourcing such a large number of judgments is not trivial as it would violate the terms-of-service of the legal research tools that would ordinarily be used to obtain them. Fortunately, UKHL judgments between 1996 and 2009 are publicly available from the UK parliament website[3]. The only problem is that cases are split across multiple web pages whose HTML format differs markedly from year to year. Therefore we created a bespoke web scraper using the BeautifulSoup library[4] which downloads the body of each case, taking into account the yearly format changes, and combining the text from successive pages into a single file.

To facilitate the potential future integration of our work with prior work on rhetorical zoning, we chose to include 69 UKHL cases previously used in the HOLJ corpus of Hachey et al. [4] in our own corpus. We then randomly selected the remaining 231 cases from the 755 cases available on the UK parliament website. The resulting files were split into individual sentences, with each sentence beginning on a new line. We first used the NLTK toolkit [5] for sentence splitting, but the complexity of a typical sentence with legal citations and abbreviations led to poor results. We then used the StanfordNLP toolkit [6] to achieve much better results. Finally, any remaining errors were manually corrected. In this way we obtained a total of 134,953 sentences in the ASMO corpus.

3.2. Annotating the Corpus

We used three experts to each annotate all 300 cases with relevant AS and MO. We had two junior annotators (both reading law at university) and a senior annotator (working as a paralegal in the UK). We provided them with a set of formal guidelines and had them participate in joint training sessions which explained the two key tasks: to identify which sentences in a case contain AS representing full agreement or acknowledgement; and to identify which judges form a conclusive MO based on those AS.

Before engaging the annotators we experimented with the BRAT [7], GATE [8] and Tagtog [9], annotation tools. But these are optimised for highlighting sub-sentence structure and not for selecting entire sentences - especially when they span multiple screen lines (as is usually the case in our corpus where sentences are 29 words long on average). It soon became clear that just trying to highlight a sentence with these tools can take longer than working out the correct label. Thus we built our own web-based annotation tool, which allows users to quickly classify UKHL sentences as *full agreement*, *acknowledgement* or *neither* and select the names of any judges forming a MO. Each expert used our tool to annotate all 300 cases over a period of two months.

[2]The ASMO corpus can be accessed at http://www.holj.ml/asmo where the full text and consensus annotations of all 300 judgements can be seen by simply clicking on the case number (or by following the direct links embedded in text of this paper discussing specific examples).

[3]See https://publications.parliament.uk/pa/ld/ldjudgmt.htm

[4]See https://www.crummy.com/software/BeautifulSoup/

3.3. Arbitrating AS Annotations

In the first task, the annotators collectively labelled 2357 sentences as containing AS. They unanimously agreed on the label of 1671 (71%) of these. Of the other 686, at least two annotators agreed in all but two instances. A consensus labelling was therefore obtained by taking the majority view where it existed, or taking the senior annotator's view in the few instances where it didn't. We found three common sources of dispute.

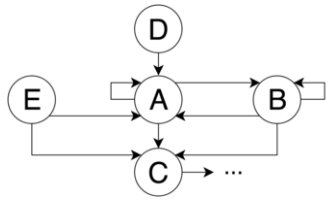
The first sort of disagreement concerns the distinction between agreement on the outcome and agreement on the reasons for the outcome. For example, one annotator mislabeled the sentence: *"In agreement with my noble and learned friend Lord Scott of Foscote, I too would allow the appeal."*, as a full agreement, when in actual fact it can only be used to infer agreement with the outcome (since, unlike sentence 1 in Table 1, it doesn't actually mention the *"reasons"*) and so it is only a partial agreement.

The second type of disagreement is on whether a sentence contains an agreement in-full with reasons leading to an outcome, or only an agreement with some subset of those reasons. For example in the sentence: *"On the basis of the wider approach to the problem of comparison which my noble and learned friend Lord Slynn of Hadley has adopted I am in full agreement with him that the rules of procedure for a claim under section 2(4) of the 1970 Act are not less favourable than those which would apply to a claim for breach of contract in the circumstances of the present cases."*, the agreement is limited to the interpretation of the 1970 Act. Hence this sentence is also a partial agreement.

The third issue concerns the ambiguous phrasing judges sometimes use to describe their agreement. For example the sentence: *"For these reasons, which really do no more than echo a part of the altogether fuller reasoning contained in the opinion of my noble and learned friend Lord Neuberger of Abbotsbury, which I have had the advantage of reading in draft and with all of which I am in complete agreement, I too would dismiss both appeals."* was understood by two experts as a full agreement with Lord Neuberger, and by one expert as a self agreement of Lady Hale together with a full agreement with Lord Neuberger. It is hard to say definitively who is correct in cases like this.

3.4. Arbitrating MO Annotations

In the second task, the experts unanimously agreed in 230 (77%) of the 300 cases. Of the other 70, at least two annotators agreed in all but six cases. Again, a consensus was obtained by taking the majority view where it existed, or the senior annotator's view in the few cases where it didn't. This time, the two main sources of dispute came from differing views about which judges some problematic AS were actually agreeing with, and from errors made in determining what MO a given set of agreements implied. A good example is shown by the graph below, depicting the full agreements in case 102:



In this case, one expert inferred {A,B,C} as the MO, presumably because (for the reasons previously explained in the last sentence of Section 2.2) it seems to be supported

by all the judges. Now, if C meant to only agree with A and B, then that would indeed be true. But, given his generic agreement, we can't rule out the possibility, unlikely as it might be, that C also meant to agree with D or E - which would then force them into the MO as well! Since we can't infer C's intentions from the AS, we must conclude along with the other two experts that there is no unambiguous MO in this case.

4. Automated Annotation Study

This section describes our approach to automating the identification of MO. To avoid over-fitting our training data, we randomly split our corpus into three equal subsets of 100 cases, called the AS-set, MO-set and AI-set, that were used in the following three subsections, respectively . The AS-set was used for the training and testing our ML model for the sub-task of AS classification. The MO-set was used for development and testing of our rules for the sub-task of MO identification. The AI-set was used to evaluate the complete AI system obtained by combining our ML model for AS classification with our rules for MO identification.

4.1. Automating AS Identification

We approached the task of AS identification as a text classification problem. Three commonly used ML algorithms, Support Vector Machines (SVM), Logistic Regression (LR) and Naive Bayes (NB) were trained to classify sentences of our corpus as acknowledgement, full agreement or neither. Each algorithm was evaluated and the F1-score was used to select the best one. All our experiments were implemented using the scikit-learn library [10] using 10-fold cross validation.

As we are only interested in a few very specific sentences from a vast corpus, our classes are highly imbalanced by nature. To prevent our models simply favouring the most frequent category of *neither* (42 525 sentences), we down-sampled the AS-Set, by randomly selecting sentences to obtain a more balanced training set with 1 292 sentences in both the *none* and *full agreement* categories, and 374 sentences in the *acknowledgement* category.

Inspired by Hachey et al.'s research on rhetorical zoning [4] and Palau et al.'s research on argumentation mining [11] we take advantage of a combination of traditional features used for the task of legal text classification. These include unigrams, part of speech tags (POS), sentence lengths, sentence position and named entities (NE). We also employ custom designed cue phrase feature, inspired by Teufel et al.'s research on rhetorical zoning [12,13]. Our features are extracted using the NLTK library [5]. The cue phrases were manually selected by a human annotator based on commonly occurring phrases in the sentences of interest. They include phrases such as: *"for these reasons", "allow/dismiss the appeal" or "I have had the advantage"*. Some of the words contained in our cue phrases (e.g. *appeal*, *dismiss*, *reasons*) are already automatically identified by the ML algorithms as the most informative unigram features. However, the cue phrases also capture their word order, making them a valuable feature. Our POS and unigram features were normalized using term frequency-inverse document frequency (TFIDF).

The best individual feature are unigrams, followed by POS tags, cue phrases, NE, position and length. The high performance of cue phrases alone suggests AS are indeed

	SVM		LR		NB	
	Ind	Cum	Ind	Cum	Ind	Cum
Unigrams	0.938	**0.938**	0.938	0.938	0.922	0.922
POS tags	0.907	**0.938**	0.893	0.935	0.851	0.916
Cue phrases	0.800	0.935	0.802	0.939	0.800	**0.936**
NE	0.490	0.935	0.490	0.942	0.186	0.931
Position	0.342	0.937	0.342	**0.943**	0.186	0.931
Length	0.282	0.937	0.329	**0.943**	0.186	0.922

Table 2. *ML experiments reporting weighted-average F-scores for 10-fold cross-validation for individual (Ind) and cumulative (Cum) performance of features.*

formulaic. On the other hand, the drop of performance between unigrams and cue phrases point to the necessity of employing a ML model to achieve F-score of 0.90 and above.

As shown in Table 2 the LR model performs the best, achieving an F-score of 0.943. It is also the only model able to successfully integrate all of the features (except sentence length) to cumulatively improve its performance. Based on this we chose to use the the LR model trained on all features but length in our complete AI system.

4.2. Automating MO Identification

Resolving AS to obtain MO consists of two sub-tasks, parsing the AS to build a graph of the case relations and resolving such graph to identify MO. To build a graph of judge agreements we parse the AS based on our observations of different types of AS structures from Section 2.1 Table 1. First however, we remove all sentences with a number in them, since these can't be AS and are a result of our ML model misannotation. The number in a sentence is a proxy for a case citation, indicating an agreement with a past judgment, or specification of a legal point of another judge, indicating only a partial agreement. We remove all acknowledgements which were not followed by a full agreement, since they are not necessary for the purpose of anaphora resolution. We remove the parts of the sentences proposing orders, since orders sometimes contain names of the judges who the agreement isn't with. We remove the acknowledging sentences if the names of the judges in them don't match the names of judges in full agreement sentences. The remaining AS sentences of a judgment are merged together. We check an indication of a self-agreement expressed by phrases such as *"For these reasons"*. Finally, we extract the names of the judges. For each case we store the agreements in a dictionary, representing our graph.

To resolve our graph we follow three rules that we believe capture the common inference patters that we previously explained in Sections 2.2 and 3.4. First we take the transitive closure of the agreement graph, using the well-known Floyd-Warshall algorithm. Intuitively this progressively adds implicit agreements from a judge A to any judge C who is agreed with by any judge B that A already agrees with. Second we take a qualified reflexive closure of the resulting graph, which means adding self agreements to any judges that have no outgoing edges (i.e. who have not expressed a full agreement with any other judge). Finally we take each judge in turn and see what precise set of judges they agree with. If any such set has the support of more than half the judges then it is taken as the MO. Applied on our MO-set, our method finds (human-identified) MO from (human-classified) AS with 89% accuracy.

4.3. Complete AI System

We built our full AI system by combining the ML model of Section 4.1 with the rules of Section 4.2. When tested on the the independent AI-set, our system replicated the consensus MO with 81% accuracy. By contrast, the average expert agreement with the consensus MO is 91%, with the lowest being 85%. Hence we claim our system is close to achieving a human level of performance. It also significantly outperforms all obvious baselines we tested such as saying there isn't any MO (29%), choosing the single most cited judge (38%), choosing the judge most mentioned in AS (48%), choosing the set of judges with opinions longer than some optimal number of sentences (43%), or choosing the single judge with the most sentences (48%).

4.4. Error Analysis

There are three reasons why we don't achieve 100% agreement with human annotators. The first cause of error is imperfectly parsed AS caused by an unusual formulation of a sentence. For example in case 233, the wording *"in the manner Lord Hutton proposes"* is used to express an agreement with an order, instead of the traditional expression which explicitly contains the word *"order"*, thereby confusing our system. To resolve these instances in the future, a statistical approach trained on a larger data set could perhaps learn to correctly identify such unusual expressions of an agreement.

The second cause of error arises when annotators arguably misannotate a case. The two common sources of annotator error in ASMO corpus are exemplified in Figure 2. The first error arises in instances of complex transitive agreement (e.g. case 90). Our experts labelled this as having no MO. But we would argue this is incorrect as there are three judges {C,D,E} all in mutual agreement once we add in the implicit transitive agreement of E with himself. The second error arises where there are only two agreements with multiple judges. For example in case 300 the human consensus is that the MO is {B,E}, but we would say there is no MO, as only two judges {A,C} actually agree on this.

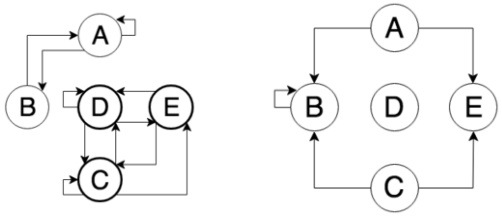


Figure 2. *Arguably mislabeled cases 90 (left diagram) and 300 (right diagram) from our corpus.*

Finally our ML model sometimes misclassifies AS. Since most of our cases have only 5 judges, a single erroneous agreement in our graph can often result in identifying the wrong MO. The low error rate we report for our ML model is therefore amplified on the task we employ it for. Particularly confusing for our model are partial agreements. For example the sentence *"I also agree with the supplementary observations made by Lord Walker and by Lord Neuberger in their opinions."* in case 91, contains the word stems *"agree"* and *"opinion"* which our ML model associates with full agreement and yet the the phrase *"supplementary observations"* gives away that this agreement is not with the full reasoning.

5. Conclusion

Our goal was to develop a computational method for identifying AS and MO in UKHL case law. To this end we demonstrated the considerable difficulties inherent in this task and we introduced a novel way of approaching these challenges from a graph-theoretic point of view. We also introduced an extensive expert-annotated ASMO corpus together with an AI system that is able to automatically identify explicit AS and implied MO with an accuracy approaching that of humans (81%).

Although we have only been concerned with the identification of MO supported by explicit inter-judge expressions of agreement, this is sufficient to establish the existence of legally binding MO in over two thirds (71%) of our corpus. Moreover, we believe the work we have done is a necessary stepping-stone towards the development of more sophisticated methods that might attempt to resolve the remaining cases through more nuanced partial agreements between judges and/or appealing to implicit semantic similarities in their legal arguments.

References

[1] J. Wilson. UKSC judgments: the case for a single, identifiable majority opinion UKSCBlog. http://ukscblog.com/uksc-judgments-the-case-for-a-single-identifiable-majority-opinion/.

[2] J. Lee. A defence of concurring speeches. *Public Law*, (Apr):305–331, 2009.

[3] B. Hale. Judgment Writing in the Supreme Court Brenda Hale UKSCBlog. http://ukscblog.com/judgment-writing-in-the-supreme-court-brenda-hale/.

[4] B. Hachey and C. Grover. Extractive summarisation of legal texts. *Artif. Intell. Law*, 14(4):305–345, December 2006.

[5] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[6] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[7] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[8] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLOS Computational Biology*, 9(2):1–16, 02 2013.

[9] J. M. Cejuela, P. McQuilton, L. Ponting, S. J. Marygold, R. Stefancsik, G. H. Millburn, B. Rost, and the FlyBase Consortium. tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles. *Database: The Journal of Biological Databases and Curation*, 2014:33, 2014.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] R. M. Palau and M. Moens. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA, 2009. ACM.

[12] S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 103–110, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[13] S. Teufel and M. Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.