Hierarchical Model Rule Based NLP for Semantic Training Representation Using Multi Level Structures

Fangmian Liu[1], Qiyuan Bian[2,*]
[1]Henan Vocational College of Tuina, Luoyang, Henan 471023
[2]Fudan University, Yangpu, Shanghai, 200438, China
E-mail: liufangmian@163.com
*Corresponding author

Keywords: NLP, hierarchical model, semantic training, multi-level structures

Recieved: October 20, 2023

During evaluation of large amounts of natural language texts, the utilisation of multi-level models is essential for the purpose of extracting knowledge that is relevant. It is essential to complete these duties to solve a variety of concerns relating to the development of textual information as well as its analysis. It is necessary to have a substantial quantity of annotated texts that contain various levels of lexical, syntactical, semantic, and narrative information to develop multi-level models for natural language texts. due syntactical annotations are maintained in a tree structure, these annotated texts are frequently referred to as text corpora or treebanks. This is due of the tree structure. Semantic treebanks are a relatively new development in this area that were introduced not too long ago. These treebanks join syntactical trees through logically-expressed smart representations of phrase sense. During the last few years, a great number of semantic treebanks that contain superficial as well as deep semantic information have been constructed. There have been a lot of different ways created, both manually and mechanically, for generating semantic treebanks. Because there aren't many standards that are universally accepted in this quickly developing subject, many semantic banks include vastly varied kinds of information. This is especially true on the lexical level. The authors of this work investigate a variety of semantic treebanks and the ways in which such treebanks could be used for text modelling. They investigate the various kinds of information, such as semantic, narrative, syntactical, and lexical data, that are stored in these treebanks. The authors also study the quantity and character of relevant corpora in addition to the key tools utilised for working with the data included within treebanks. These methods have a wide range of applications in decision-making processes that are concerned with the generation and analysis of text. An example of their usage is for annotating and retrieving information resources to facilitate collaborative development of a domain information space based on ontology, particularly in scientific research and learning. Additionally, you can use them to create and re-write texts for a variety of purposes, including fiction writing, marketing, and scientific communication.

Povzetek: Raziskava obravnava razvoj in analizo tekstovnih informacij s pomočjo hierarhičnih NLP modelov, ki uporabljajo večnivojske strukture za semantično usposabljanje, zlasti z uporabo semantičnih drevesnih struktur.

1 Introduction

The use of multi-level modeling for natural language texts is beneficial for solving various problems related to text generation, analysis, annotation, and retrieval. To create reliable multi-level models, a significant amount of annotated textual data must be analyzed. Treebanks, which are annotated text corpora, are a useful resource for modeling purposes. Text analysis requires different types of input data depending on the dimensions being analyzed. There are several significant classes of modeling data, including semantic data, which represents the text's direct meaning and is often represented through first-order logic. Narrative data captures the narrative elements of the text, including the genre, intended audience, and author's style. Syntactical information describes sentence structure andfunctions which can be stored using constituency and dependency structure trees. Lastly, lexical information includes specific word details, such as part-of-speech describing, lemmas, data.

The objective of modeling the lexical aspect of text is to provide guidance for selecting the appropriate phrase to

convey deliberate meaning. This involves picking a synsetbeginning a list of alternative expression and determining level of evidence for the synset. To create a robust model for word selection, it is necessary to consider lexical, syntactical, and narrative information. Labeling the meanings of words, constructing dependency trees for phrases and clauses, and analyzing description components all contribute to this process. Synsets rely on semantic relations between synsets, especially hyponymy, as opposed to enhancing predicative semantic restraints, which characterise only a small subset of WordNet relations [19-21]. This recursive approach to language allows for the definition of any constraints on meaning.Another attempt to establish semantic roles focuses exclusively on verbs; examples include the sense id models VerbNet [22] and PropBank [23]. Dependency trees are important for modeling natural language texts because the choice of a dependent word may be influenced by the head word in the dependency link. In addition to lexical information, narrative data plays a crucial role because the selection of a restrictedalternative expression can be induced by numerouselements, such as text's genre, background of narrator, and intended audience. To aid decision-makers in choosing the appropriate type of sentence, the number and type of clauses, and other structural choices, modeling the syntactical dimension of text is also essential. This requires considering both the narrative information that affects sentence length and structure, as well as the semantic depiction of sentence's content in verb phrase logic system. Syntactical data can be represented using constituency trees, which can be constructed from dependency trees.

The semantic component of text modeling involves organizing a text fragment to convey information effectively. This involves determining the appropriate length of the text fragment, the level of detail required, how to organize the content, and adding specific details to important information. It also involves determining the number of sentences and their content to create a coherent narrative structure. To model this level, predicate logic semantic information, high-level syntactical information (such as the top of dependency and constituency trees), and detailed narrative annotation, including text fragments and their components, mode of narration, narrative goals and links, and other relevant information, are necessary. The advancement of NLP technology in recent years has been remarkable, with the BERT paper by Devlin et al. [24] being a pivotal moment that introduced a new neural network architecture and training method with a significant impact on the expansion of NLP. BERT is a highly versatile tool for a wide range of NLP applications, improving the performance of many benchmarks by over 20%. The BERT neural network is based on self-attention, where the algorithm infers a hidden word from its left and right context during training by focusing on each surrounding word. A trained BERT model, also known as a Masked Language Model, can perform various NLP tasks, including paraphrase extraction, question answering, and semantic similarity testing. However, BERT has some limitations, such as its computational complexity, which is proportional to O(N2), where N is the dimension of the hidden layer and FLOP is the number of floating-point operations. As a result, input sequence lengths are typically limited to 512 or 1024 tokens.

When existing queries produce unsatisfactory results, business users can modify their search queries by selecting more specific or general updated concepts. Adding new terms to existing business taxonomies to better reflect the dynamic world news poses two challenges. Firstly, locating existing BI-specific datasets and vocabularies to enrich taxonomies, and secondly, adding new concepts while preserving the current taxonomy's respect for how business concepts are structured.

2 Literature review

According to the International Monetary Fund [18], private tax rulings, also known as PTRs, are a form of guidance that taxpayers can request from tax authorities in order to gain a better understanding of how tax rules apply to their particular circumstances. When taxpayers rely on PTRs, they are often shielded from further taxes, penalties, and interest, and the tax authority is obligated to follow the ruling. Additionally, when taxpayers rely on PTRs, the tax authority is required to obey the ruling. PTRs, on the other hand, almost always exclusively benefit the person who requests them and do not set a precedent for subsequent taxpayers. Both taxpayers and tax authorities benefit from increased consistency and clarity in the administration of tax legislation because to the existence of the private tax ruling system. .For intricate or unusual economic transactions, taxpayers can submit tailored PTR applications. To improve transparency and predictability in tax systems, the IMF suggests disclosing private rulings with appropriate redactions.

Private tax rulings are a useful instrument for reducing or even getting rid of the tax risks that are involved with big commercial transactions. They frequently serve as the basis for subsequent interpretations and reveal the initial attitude of tax officials in the area. While private tax rulings are frequently used by tax planners and advisors for large corporations, smaller taxpayers also use them as a safeguard. However, preparing a request for a tax ruling is often too complex for the average taxpayer, which is why requests are usually submitted by tax advisors or lawyers and require significant effort from highly skilled tax professionals. Nonetheless, obtaining a private tax ruling is generally safer, less expensive, and quicker than litigating taxes in court.

Taxpayers do not submit requests for a judgement carelessly. They typically only perform it in complicated and uncommon instances. The body of tax judgements provides a direct glimpse into taxpayers' everyday issues and illuminates previously hidden patterns of behaviour. As a result, policymakers can benefit from quantitative analysis of the corpus since it identifies problematic regions that could be resolved by modifying the tax code.

Table 1: Comparison of state of art models

Ref	Technology	Challenges
[18]	Private Tax Rulings	Often shielded from further taxes, penalties
[19]	WordNet relations	Opposed to enhancing predicative semantic restraints
[20]	WordNet relations	Rely on semantic relations
[21]	WordNet relations	Focuses exclusively on verbs
[22]	VerbNet models	Difficulty of extending a manually-curated resource
[23]	PropBank	Takes a practical approach to semantic representation, adding a layer of predicate-argument information
[24]	BERT	Expensive and requires more computation because of its size.

The Table 1 proves and illustrates the drawbacks of the existing state of the art models and its computational complexity as a major challenge of the domain. Relevant knowledge must be extracted through the analysis of vast volumes of natural language texts and the application of multi-level models. Completing these tasks is necessary in order to address a number of issues regarding the creation and interpretation of textual data. To create multi-level models for texts in natural language, a significant amount of annotated texts with different levels of lexical, syntactical, semantic, and narrative information are required. Text corpora or treebanks are common terms used to describe these annotated texts that have syntactical annotations kept in a tree structure.

3 Proposed work

A measure of the degree to which two sections of text are semantically same is referred to as the "Semantic Textual Similarity" (STS). Rather than giving a straightforward yes or no answer, algorithms that evaluate semantic similarity typically produce a ranking or percentage that indicates the extent of textual similarity rather than a simple yes or no answer. Unfortunately, there is no definition of semantic equivalence that is globally acknowledged, which means that there is no definition of STS that is either universally accepted or widely accepted. In order to properly evaluate semantic equivalence and similarity, it is essential to take into account the context in which a word or phrase is employed. The context in which a word or phrase is used is what establishes its meaning; this, in turn, defines how semantically similar it is to other words or phrases. Over the years, various algorithms and techniques have been developed for measuring semantic (textual) similarity, including knowledge-based, corpus-based, and deep learning approaches. By leveraging large corpora and deep learning, semantic similarity techniques are able to quantify the semantic similarity between phrases. The "distributional hypothesis," which assumes that "similar words frequently co-occur," forms the basis of these techniques, but does not consider actual suggesting of words.

The application of transformers-based deep neural network techniques has shown higher performance when compared to the majority of traditional approaches, and the recent success of these techniques has completely reshaped the field of semantic similarity. Devlin and colleagues [3] conducted a ground-breaking study in which they presented a novel neural network as well as a new training approach, which they combined referred to as BERT. The natural language processing (NLP) algorithm known as BERT, as well as its several variants, is regarded as one of the most effective algorithms currently on the market, according to a number of studies and benchmarks [29–34]. By analysing both the left and right contexts in which a word or phrase appears, language models that are based on BERT are able to discern between several meanings of the same word or

phrase. The proposed project's workflow is depicted in figure 1, which may be found here.

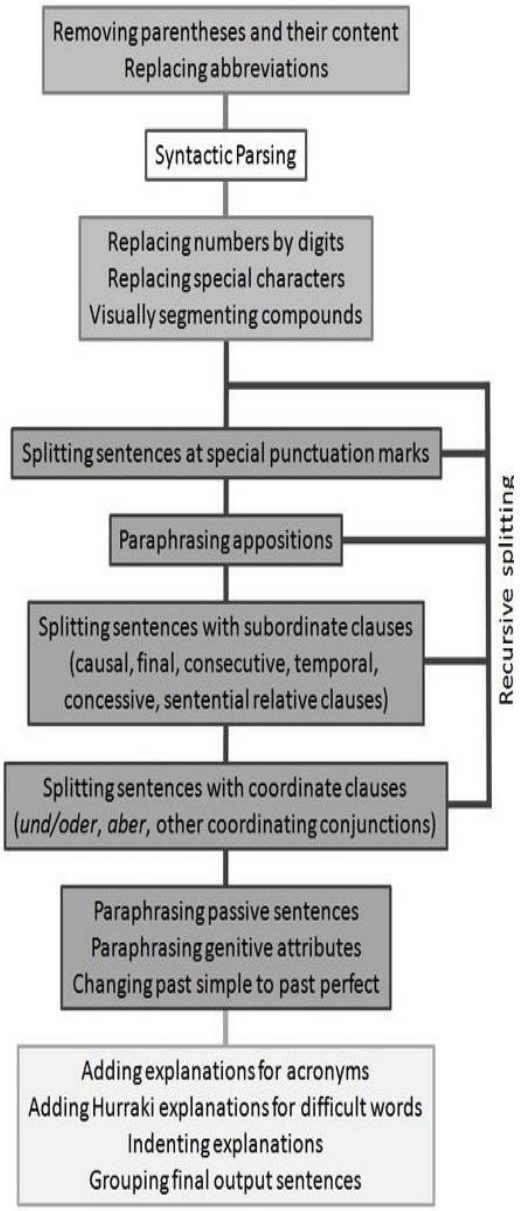

Figure 1: Flow process of NLP in semantic training representation

Over the course of centuries, legal career has progressed its own specialised language, commonly known as legal (sub)language, which lawyers use to discuss the law. The foundation of this language is legislative language, which refers to the terminology used to draft legislation. Linguists often classify legal language. A sublanguage has its own unique grammar, a narrow scope [35], specific lexical [36], syntactic [37], and semantic constraints [38], and allows for 'deviant' grammatical rules that are not allowed in the dominant language. Given that the legal profession has been dubbed "a profession of words," it is crucial to have a strong command of legal language. We will not delve into the intricacies of legal jargon here; those who are interested can find a plethora of literature on the topic.

The legal profession has developed its own specialised language over the course of several centuries, and this language, which is frequently referred to as legal (sub) language, is the language that lawyers use to explain the law to their clients. Most of this language is comprised of legislative language, which is the terminology that is used while writing legislation. Linguists frequently consider legal language to be its own distinct sublanguage [39]. A sublanguage is characterised by its unique syntax, restricted usage, the imposition of lexical, syntactic, and semantic limitations, and the presence of 'deviant' grammar rules—that is, rules that are not permitted in the grammar of the dominant language. Since the legal profession is often referred to as "a profession of words," having a strong grasp of legal language is crucial. There is an abundance of literature available for those interested in exploring legal jargon, so we will not delve into the specifics here.

3.1 NLP – For hierarchical modeling

Processing legal content presents a problem for natural language processing systems. In order to get the best results possible from NLP algorithms, they should be trained on vast corpora of plain text. A good example of this kind of corpus is the Open Super-large Crawled Aggregated (OSCAR) corpus, which was only recently made public and has a size of many terabytes. For the purpose of processing legal language, ordinary text language models (LM) need to be modified to legal language, ideally using a large corpus of previously processed legal texts. However, it is extremely difficult to acquire legal corpora, and many of them are kept privately, making them inaccessible to academic scholars. The scholarly works, statutory instruments, judicial decisions, court memoranda and pleadings, commentary on statutes, and administrative law decisions would all be included in a comprehensive legal corpus. In our study, we utilized the corpus of private tax judgements to enhance a pre-existing BERT model for Polish, as no corpus of this kind is currently available.

3.2 Similarity among the legal semantic text

In the subject of law, a straightforward semantic or linguistic similarity comparison might not be enough. Legal scholars look for content similarities within a legal framework, such as a legislation, rule, or judgement, where the words and phrases must have the same legal meaning in order for the

similarities to be considered relevant. The fact that two different sections of text contain the same terms is not sufficient. Legal Semantic Text Similarity (LSTS) is a measure of legal and semantic textual similarity between two segments of text, always evaluated in the same legal context. This measure compares the legal and semantic textual equivalence of the two segments of text. We want to stress that LSTS is only applicable to textual similarity and that an LSTS algorithm is unable to identify the relevance of a textual similarity. In other words, the system that conducts the evaluation of legal semantic similarity needs to be able to differentiate between the legal meanings of a word or phrase when it is used in a variety of distinct legal situations. When it comes to matters pertaining to taxes, the legal framework is often determined by the statute or legislative instrument to which the terminology in question applies. It can be difficult to ascertain what a particular word, phrase, or paragraph in a document is supposed to signify, and the ability to discriminate between the statutory and common meaning is essential in a great number of financial and legal disputes. We suggest a fresh approach to finding tax judgements with legal semantic textual similarities. Finding a way to identify judgements that are semantically similar is our goal, as one ruling may be a tens of page legal language. Experiments revealed that the tax authority stance, which mentions pertinent provisions of statutes, other tax judgements, court decisions, and other sources, is the most helpful element of the ruling for semantic similarity searches.

This section, which utilizes formal legal terminology, offers, in general, a summary of Sections 2, 3, and 4 of the decision (circumstances, enquiries, and the taxpayer's legal standing). In our experiments, we determine the degree to which the various aspects of the position taken by the tax authorities share a semantic similarity. We are aware that the section on the tax authority provides further information regarding the specific legal setting of the ruling. The clustering, cosine similarity, and SBERT vector embeddings that we use form the foundation of our methodology. We compute the semantic textual similarity of SBERT sentence vector embeddings. The SBERT sentence vector embeddings are used. We determine the cosine similarity of the selected vector to other vectors by conducting an analysis on the vector embeddings of each and every component of each and every PTR. The cosine similarity between the RRNN's two vectors, uu and vv, is defined by Equation 1, which may be found below.

$$similatiry\ (u, v) = \frac{|U||V|^T}{\|U\|\|V\|} \tag{1}$$

Using cosine similarity, we identify the K tax authority positions that are the K closest neighbours in the embedding space. The embedding vectors are then projected using UMAP, and the HDBSCAN clustering algorithm is then applied to the UMAP output. Using this strategy, we can identify discrete clusters that accurately reflect the fine semantical characteristics of the judgements.

4 Experimental analysis

The implementation of our specialized feature extraction methodology is the first step in the processing of the PTR corpus. This step takes place in the beginning. The pipeline compiles a comprehensive database of information pertaining to each judgement that it processes. This includes citations to previous judicial decisions, legislation, rules, and schedules, as well as references to other PTRs, rules, and schedules. Additionally, the pipeline extracts references to other relevant PTRs. Additionally, the document is broken up into paragraphs and sentences by the pipeline, which makes use of the Polish translation of Stanford's Stanza. The wording of the PTR is broken up into four sections: an explanation of the circumstances (the facts), the taxpayer's queries, the taxpayer's legal position, and an explanation of the perspective of the tax authority.

Next, we independently build sentence BERT (SBERT) vector embeddings for each part using our improved BERT model, which is derived from Polish BERT [6]. In the event that a section of text has more than 512 BERT tokens, we break it up into subsections comprised of sentences and paragraphs. When it is possible, one sentence from the preceding paragraph will be carried over into the next paragraph. If the vector embedding of a PTR component creates more than one vector, we follow the established protocol and take the mean of all of the vectors into account. In our research, we employ a number of well-known, open-source Python packages:

- Tokenizers, sentence-transformers, and hugging face transformers packages
- HDBSCAN with UMAP Scikit-learn

The following is the technique that should be followed in order to locate PTRs with comparable semantic features. We start by selecting a PTR of interest from the precomputed data frame to use as the reference PTR, and then we extract its SBERT embedding. We compare the reference PTR with each and every other PTR utilising the cosine similarity metric, and then select the top K nearest neighbors. The value of K will be set to 500 for the sake of our experiment. The cosine similarity of the three PTR sections' 500 nearest neighbors is depicted in Figure 2, which may be found here. In order to find the PTRs with the highest similarity

coefficients, the tax authority position of the reference PTR and the top k PTRs are compared. For the next stage of processing, we select the list of 500 PTRs whose tax authority parts have the most striking similarities to those found in the reference PTR.

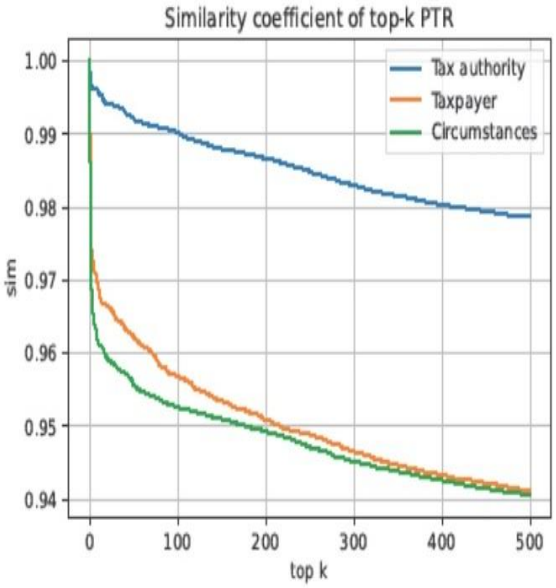

Figure 2: Similarity Index of 500 neighbor to the reference PTR

A tax specialist must evaluate the list of the 500 PTRs that are the most identical. Textual similarity alone does not establish legal resemblance between a PTR and the reference PTR. The certified tax advisor who carried out this manual analysis is familiar with the issue mentioned.

Analysis's outcome was excellent—indeed, it was surprising—because all 500 PTRs were indeed validly comparable. Related outcomes were obtained in experiments using other reference PTRs. The list makes it clear that the system discovered PTRs that were related in both syntactic and semantic terms. Four instances of question-related PTRs that the algorithm identified are provided in Table 2. Due to the fact that translation might remove some syntactic and semantic elements, we chose not to translate the question wording. However, readers who are not Polish might notice that the questions are phrased differently and have a distinct syntax.

All of the mentioned decisions are related to the same matter and are identical in terms of their subject matter and legal standing. It is important to note that while there is a small degree of word and phrase similarity among taxpayers' responses, there is a very high degree of cosine similarity.

4.1 Virtualization

The SBERT vector embeddings were used to conduct an analysis on the 500 PTRs that shared the highest degree of similarity. Because the dimensions of the embedding vector space (N=768) is too high for direct viewing, we make use of the UMAP technique to project embedding vectors into a space that is either two or three dimensional in order to make it easier to visualize the data. A technique for reducing the number of dimensions that is known as UMAP, which stands for uniform manifold approximation and projection for dimension reduction, can be used to depict high-dimensional vectors. The results of plotting the projections are illustrated in figure 3, which may be found below. The two-dimensional diagram already gives hints that the embedded vectors cluster together, and the three-dimensional figure proves what the two-dimensional diagram already suggests.

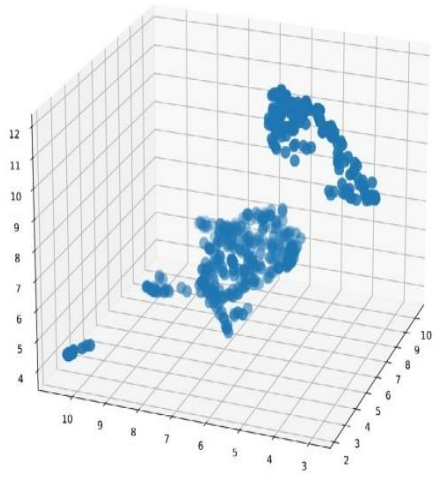

Figure3: Plots of the 500 most comparable decisions' three-dimensional embedding vectors

We used a clustering technique since the visualisations show that vector embeddings include clusters. We conducted the popular HDBSCAN [9] algorithm on a 2-dimensional UMAP projection using the Scikit-learn toolkit. The same outcomes were obtained when HDBSCAN was applied to vector embeddings, although the runtime was significantly longer.

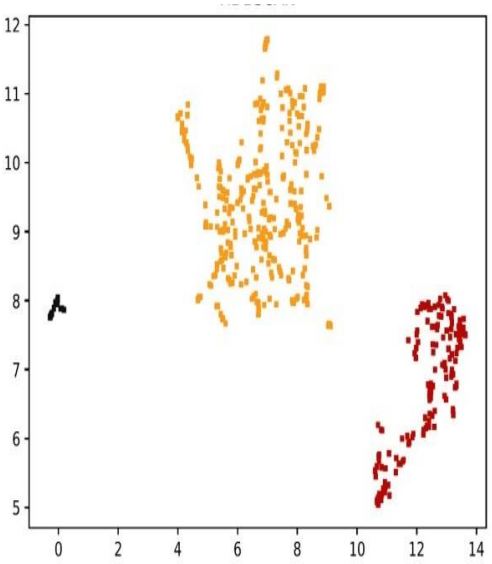

Figure 4: Plot of detected clusters, coded by color

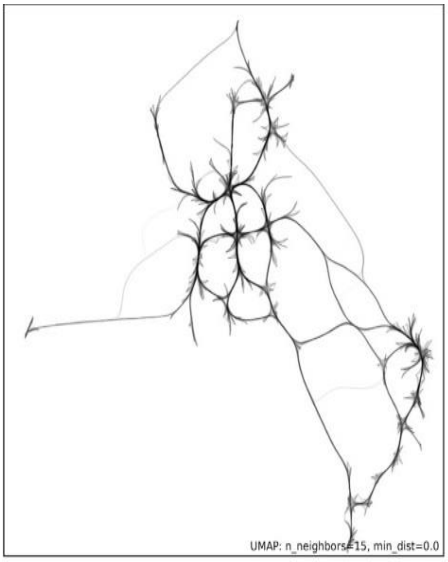

Figure 5: Plot of the internal structure of clusters with edge bundling

Three clusters were present, as evidenced by the clustering algorithm's output (Fig. 6). The "hammer" plot, which is second plot in Fig. 6, displays inner organisation of clusters with edge shoving. For more information about this plot, check the HDBSCAN documentation. Red ovals in Fig. 6's dendrogram of samples (rulings) denote clusters [40].

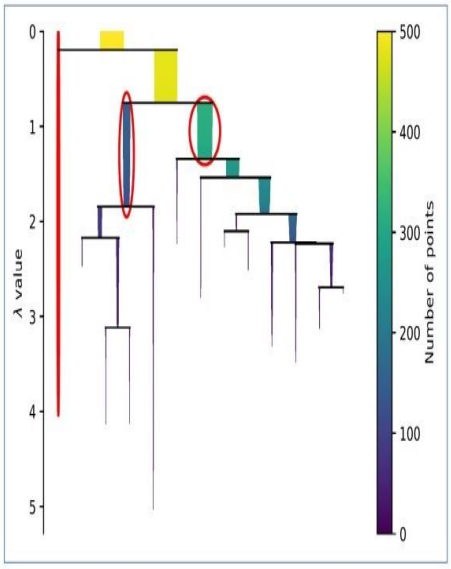

Figure 6: Dendrogram cluster plot

4.2 Cluster analysis

Interestingly, we observed the formation of clusters containing obviously distinct embedding vectors. Upon examination, we discovered that judgements within a cluster share nontrivial semantic properties. The largest cluster we identified consists of decisions addressing more general issues, such as the applicable VAT rate [41] that landlords should use when billing tenants for the use of various services, including water, gas, electricity, sewage, and so on. The second, more concentrated cluster is made up of decisions addressing a more specific question, such as the VAT rate landlords should apply when billing tenants for electricity consumption..

The cluster featured unusual judgements, such as special situations and agreements and judgements that were decided in a manner that deviated from accepted practise (case law). We also looked at two other examples: real estate sales and tax judgements pertaining to the IP BOX tax structure [42]. Cluster searches in both situations revealed distinct clusters. It is amazing how well-kept the semantic information is in even the averaged embedding vectors. There is unquestionably more to learn and learn about in this area. We conclude that the discovery of semantically and legally related judgements as well as the homogenous grouping of the rulings is facilitated by our two-step search method, which entails discovering the K nearest neighbours (K most similar rulings) using the cosine similarity metric.

5 Conclusion

This study presents a method for extracting thematic characteristics from news articles using corpus-based thematic characteristics extraction from news articles, pre-trained word embeddings, linked open data, and lexical datasets. Approximately 91% of taxonomy concepts were present in selected datasets, and their corresponding semantic data was retrieved for taxonomy enrichment. By adjusting the cosine similarity threshold while selecting relevant ideas, the enhanced depth of the business taxonomy can be altered. The scope of a taxonomy with a high threshold is more limited. However, a low similarity threshold may permit the extraction of meaningless concepts. This essay offers original contributions in two areas namely, a way to look for decisions that are most like a source decision, and locating groups among set of decisions that are most comparable. The findings of this manuscript suggest a number of future research areas, including, semantic similarity in law research and studying the composition of similar orzecze clusters.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest

Funding Statement

This study did not receive any funding in any form.

References

[1] Abeille´,A.,2012.Tree banks: Building and Using Parsed Corpora. Text, Speech and Language Technology, *Springe*r Netherlands.

[2] Abend, O., Rappoport, A., 2013a.Ucca: A semantics-based grammatical annotation scheme, in: *IWCS*.

[3] Abend, O., Rappoport, A., 2013b.Universal conceptual cognitive annotation (UCCA), in: *ACL*.

[4] Anikin, A., Litovkin, D., Kultsova, M., Sarkisova, E., 2016. Ontology-based collaborative development of domain information space forlearning and scientific research, in: Ngonga Ngomo, A.C., Krˇemen, P. (Eds.), Knowledge Engineering and Semantic Web: 7th International Conference, KESW 2016, Prague, Czech Republic, September 21-23, 2016, *Proceedings*, pp. 301–315.

[5] Anikin,A.,Sychev,O.,Gurtovoy,V.,2019.Multi-level modeling of structural elements of natural language texts and its applications. *Advances in Intelligent Systems and Computing* 848,1–8.

[6] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N., 2013. Abstract meaning representation for sem banking, in: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for *Computational Linguistics*. pp. 178–186.

[7] Bonial, C., Bonn, J., Conger, K., Hwang, J.D., Palmer, M., 2014. Propbank: Semantics of new predicate types, in: Proceedings of the NinthInternational Conference on Language Resources and Evaluation (LREC-2014), *European Language Resources Association* (ELRA).

[8] Bos, J., 2011. A surveyofcomputationalsemantics:Representation,inferenceand knowledge in wide-coveragetextunderstanding. *Language and Linguistics Compass* 5, 336–366.URL:

[9] Bos, J., Basile, V., Evang, K., Venhuizen, N., Bjerva, J., 2017. The groningen meaning bank, in: Ide, N., Pustejovsky, J. (Eds.), Handbook of Linguistic Annotation. *Springer*. volume 2, pp. 463–496.

[10] Butler,A.,2017.Tree bank semantics parsed corpus. Carnie, A., 2013. Syntax: A Generative Introduction. Introducing Linguistics, *Wiley*.

[11] Dixon,R.,2009.Basic Linguistic Theory Volume 1: Methodology. Basic Linguistic Theory, *OUP Oxford*.

[12] D Prabakaran, S Sriuppili, “Speech Processing: MFCC Based Feature Extraction Techniques-An Investigation*”, Journal of Physics: Conference Series*, Vol. 1717, No. 1, Pp. 1-7, 2021.

[13] Dixon,R.,2012. Basic Linguistic Theory Volume 3: Further Grammatical Topics. Basic Linguistic Theory, *OUP Oxford.*

[14] Fellbaum, C.(Ed.),1998.Word Net: an electronic lexical database. *MIT Press.*

[15] Grimm, S., Hitzler, P., Abecker, A., 2007. Knowledge representation and ontologies, in: Studer, R., Grimm, S., Abecker, A. (Eds.), SemanticWeb Services: Concepts, Technologies, and Applications. *Springer Berlin Heidelberg, Berlin, Heidelberg*, pp. 51–105

[16] Hershcovich, D., Abend, O., Rappoport, A., 2017. A

transition-based directed acyclic graph parser for ucca, in: *Proc. of ACL*, pp. 1127–1138.

[17] Kamp, H., Reyle, U., 1993.From Discourse to Logic. Introduction to Model theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. *Kluwer, Dordrecht*.

[18] Kultsova, M., Anikin, A., Zhukova, I., 2015.Ontology-based method of electronic learning resources retrieval and integration, in: 2015 6th International Conference on *Information, Intelligence, Systems and Applications* (IISA), pp.1–6.

[19] McDonald, R., Crammer, K., Pereira, F., 2005.Online large-margin training of dependency parsers, in: Proceedings of the 43rd AnnualMeeting on Association for Computational Linguistics, Association for Computational Linguistics, *Stroudsburg*, PA, USA. pp. 91–98.

[20] Palmer, M., Gildea, D., Kingsbury, P., 2005.The proposition bank: An annotated corpus of semantic roles.*Computational Linguistics* 31.

[21] Parsons, P., Parsons, T., 1990. Events in the Semantics of English: A Study in Subatomic Semantics. Current studies in linguistics series, *MITPress*.

[22] Ruppenhofer,J., Ellsworth,M., Petruck,M., Johnson,C., Scheffczyk, J.,2016. Frame Net II: Extended Theory and Practice. *Institut fu¨r Deutsche Sprache*, Bibliothek.

[23] Schuler, K.K., 2005. Verbnet: A Broad-coverage, Comprehensive Verb Lexicon. Ph.D.thesis. Philadelphia, PA, USA. AAI3179808.

[24] Taylor, A., Marcus, M., Santorini, B., 2003. The penn tree bank: An overview, in: Treebanks. Text, Speech and Language Technology, vol 20..*Springer*.

[25] D.Prabakaran and H.Sathyapriya, "A Review on Methodologies and Performance Analysis of Device Identity Masking Techniques", *International Journal of Scientific & Technology Research*, Vol. 8, No. 12, Pp. 2018-2022, 2019.

[26] De Martino, Graziella, Pio Gianvito, Ceci Michelangelo. (2021) "PRILJ: an efficient two-step method based on embedding and clustering for the identification of regularities in legal case judgments." *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-021-09297-1.

[27] Devlin Jacob, Chang Ming - wei, Lee Kenton, Toutanova Kristina (2019) "BERT: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171– 4186.

[28] Dhivya Chandrasekaran and Vijay Mago. (2021) "Evolution of Semantic Similarity—A Survey." ACM Comput. Surv. 54, 2, Article 41 (March 2022), 37 pages. DOI:https://doi.org/10.1145/3440755.

[29] IRS (2008) "The Complexity of the Tax Code" Taxpayer Advocate Service — 2008 Annual Report to Congress — Volume One, Internal Revenue Service, Washington, DC.

[30] D. Prabakaran and R. Shyamala, "A Review On Performance Of Voice Feature Extraction Techniques," 2019 3rd *International Conference on Computing and Communications Technologies* (ICCCT), Chennai, India, 2019, pp. 221-231.

[31] Kumar Ankit, Makhija Piyush, Gupta Anuj. (2020) "Noisy text data: Achilles' heel of BERT". Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), pp 16–21.

[32] Leland McInnes, John Healy, James Melville. (2018) "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction", *ArXiv e-prints* 1802.03426, 2018

[33] Leleand McInnes, John Healy, Steve Astels. (2017) "hdbscan: Hierarchical density based clustering*". Journal of Open Source Software, The Open Journal, volume* 2, number 11. 2017.

[34] Mandal Arpan, Chaki Taktim, Saha Sarbajit, Ghosh Kripabandhu, Pal Arindam, Ghosh Saptarshi, (2017) "Measuring similarity among legal court case documents" Proceedings of the 10th Annual ACM India Compute Conference, *Association for Computing Machiner*y, Compute '17, pp 1–9.

[35] Mandal Arpan, Chaki Taktim, Ghosh Kripabandhu, Ghosh Saptarshi, Mandal Sekhar. (2021) "Unsupervised approaches for measuring textual similarity between legal court case reports". *Artificial Intelligence and Law volume* 29, pp 417–451. 2021.

[36] Mellinkoff David (1963) "The Language of the Law." Little, Brown and Co. 1963 *pp. xiv*, 454.

[37] Reimers, Nils, and Iryna Gurevych. (2019) "Sentence-bert: Sentence embeddings using

siamesebert-networks." *arXiv preprint arXiv:*1908.10084.

[38] Ricardo J. G. B. Campello, Davoud Moulavi & Joerg Sander. (2013) "Density-Based Clustering Based on Hierarchical Density Estimates.", In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science(), vol 7819. *Springer*, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14.

[39] Shao Yunqiu, Mao Iaxin, Liu Yiqun, Ma Weizhi, Satoh Ken, Zhang Min, Ma Shaoping. (2020) "BERT-PLI: Modeling paragraph-level interactions for legal case retrieval." Proceedings of the twenty-ninth *international joint conference on artificial intelligence*, IJCAI-20, pp 3501–3507.

[40] Strąk Tomasz, Tuszyński Michał. (2020) "Quantitative analysis of a private tax rulings corpus.". *Procedia Computer Science* 2020.

[41] Waerzeggers, Christophe and Cory Hillier (2016) "Introducing An Advance Tax Ruling (ATR) Regime" Tax Law IMF Technical Note 2016 (2) (*International Monetary Fund, Washington*, DC, 20016).