Utilization of Underlying Semantic Information in Textual Data

András Kicsi

Department of Software Engineering
University of Szeged

Szeged, 2022

Supervisor:

Dr. László Vidács

Summary of the Ph.D. Thesis

University of Szeged
Doctoral School of Computer Science

Introduction

Humanity's primary channel of information is verbal or written natural language. Throughout history, written words have aided scientific endeavors, contributed to the education of the masses, have made and broken regimes and led to our current society. Even though the information is primarily stored as binary data nowadays, it still has to be transformed to natural language interpretable for human readers. The thesis deals with the extraction of the information within natural language text of various domains. Two of the thesis points are strongly connected to software development, where the source code still has a great amount of natural language with discernable semantic information, while the third thesis point deals with the automatic understanding of radiologic reports. Text is omnipresent in our everyday lives, and its proper automatic processing has immense potential. Let us address some of the most important concepts in a really brief overview first.

For a long time, humanity has been obsessed with thinking machines. Computers serve us is every second of our daily life with highly advanced capabilities that were unimaginable just a few years ago. **Artificial intelligence (AI)** has proven to be able to outperform the human brain at many tasks. Long before the currently fashionable artificial neural networks' debut, there have already been highly advanced heuristic and machine learning algorithms capable of incredible feats. Even these, however, were highly reliant on quality data.

Natural languages are the languages that are developed naturally in use. Natural languages differ from formal ones as they tend to have more exceptions, irregularities, and as a consequence, complexity. **Natural language processing (NLP)** is the process of analysing the natural language text with various goals. NLP is widely used in several domains, and recent advances really highlight its capabilities. AI assistants can understand our speech without difficulty, and chatbots are capable of expressing human-like behaviour near-flawlessly. Natural language conveys semantic meaning, and thus its processing can bridge some structural boundaries set by domain-specific restrictions such as source code syntax. One specific field of NLP is called information retrieval (IR), which deals with the extraction of valuable information from elements of a text. Our current topics all revolve around information retrieval in one way or another.

Latent Semantic Indexing (LSI) [3] is an older technique that has been used as mainstream in many tasks of semantic analysis. It relies on semantic information of text handled as vectors and produces a more compact form of vectors that results in the semantically more similar documents obtaining less distance in their vectors. LSI uses singular value decomposition to achieve this task.

Doc2Vec was introduced by Google's developers [22] and can be considered an extension of Word2Vec, a word embedding technique. It operates with vector representations of words that are transformed to a lower number of dimensions via neural networks. The hidden layer has fewer neurons than the input and output layers, and the weights of the hidden layer provide the word embedding output we need. Thus, similarly to LSI, a more compact representation is constructed. Doc2Vec differs from Word2Vec only in also adding a unique identifier for each document to the input layer, thus distinguishing different documents.

Long Short-Term Memory (LSTM) [5] is an artificial neural network used in artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. This makes it possible for the LSTM to process entire data sequences. **BiLSTM** is the bi-directional version of this model. This consists of two LSTMs: one taking the input in a forward direction and the other in a backwards direction. BiLSTMs increase the amount of information available to the network, improving the context available to the algorithm.

Conditional Random Fields (CRF) [21] is a model suitable for sequence learning, which also provides a solution to the label bias problem. CRF estimates the conditional probability of a sequence.

Bidirectional encoder representations from transformers (BERT) [4] is a transformer-based model that applies the popular attention mechanism to textual context. The model takes into account the words that occur before and after the tokens when representing them. The model can be fine-tuned by adding a single output layer, thus achieving state-of-the-art results on several natural language processing tasks. More specific tasks (downstream tasks) are built into the process, thus resulting in separate fine-tuned models. As a result, the big advantage of BERT is that the difference between the architecture of the pre-trained and the downstream model is minimal. Depending on its size, BERT contains different amounts of encoder layers and a bidirectional self-attention head. Simply put, the architecture of BERT is a set of transformer encoding layers stacked on top of each other.

Static source code analysis is a software engineering term. Software code is text structured according to the particular syntax of the language used. The analysis of source code holds a great number of obvious benefits like quality assurance, better compilers, and advanced coding practices. It has been practised and researched for many years. In contrast to dynamic analysis, static analysis does not require the software to be actually run during its examination. Such analysis can involve the construction of an **abstract syntax tree (AST)**, or on an even higher abstraction level, an **abstract semantic graph (ASG)**. By adding function calls as edges, **call graphs (CG)** can also be constructed. These artefacts serve as potent tools for analysis and play at least some part in many of our current topics.

There are some **universal metrics** our results will be displayed with. Let us discuss them here briefly.

Precision is the proportion between correctly detected or retrieved results (relevantResults) and all detected or retrieved results (retrievedResults). It computes as

$$precision = \frac{relevantResults \cap retrievedResults}{retrievedResults}$$

It basically describes what proportion of our results (retrieved tests for traceability or tokens classified as disorders for radiology understanding, for example) was correct.

Recall is the proportion between the correctly detected or retrieved results and all the results that should have been detected or retrieved. It computes as

$$recall = \frac{relevantResults \cap retrievedResults}{relevantResults}$$

It basically describes what proportion of the real results (retrieved tests for traceability or tokens classified as disorders for radiology understanding, for example) was indeed detected or retrieved.

F1-score (or just F-score in our context) is a measure that combines precision and recall is the harmonic mean. This is usually a good indicator of how well a method performs.

Accuracy describes how many of the decisions were on point. In a binary decision case, it factors in true positive (TP), false positive (FP), true negative (TN), and false negative (FN) decisions and computes as

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

In a multi-label classification case, accuracy is computed as

$$accuracy = \frac{CorrectClassifications}{AllClassifications}$$

I Feature-Extraction in 4GL Systems

The first thesis point focuses on feature extraction during software product line adoption over several software variants. The adoption process requires developers and domain experts to work strongly together on the new architecture, as knowledge of the system's features can be invaluable. The features are implemented through parts of the software code. Feature extraction attempts to extract these for each feature for easier modification or merging with the architecture. Working with an industrial partner to aid this extraction on 19 variants of a pharmaceutical logistics system presented a couple of challenges. The software was written in a fourth generation language (4GL), Magic XPA. The code of the variants was assembled through a development environment without actual coding, and the structure of the language is also significantly different from mainstream programming languages that have ready solutions for aiding feature extraction. Magic applications consist of programs and their subtasks calling each other, and also logical and data units. We have introduced several new approaches to the feature extraction topic through our work, aiming to provide more useful information for developers and domain experts alike.

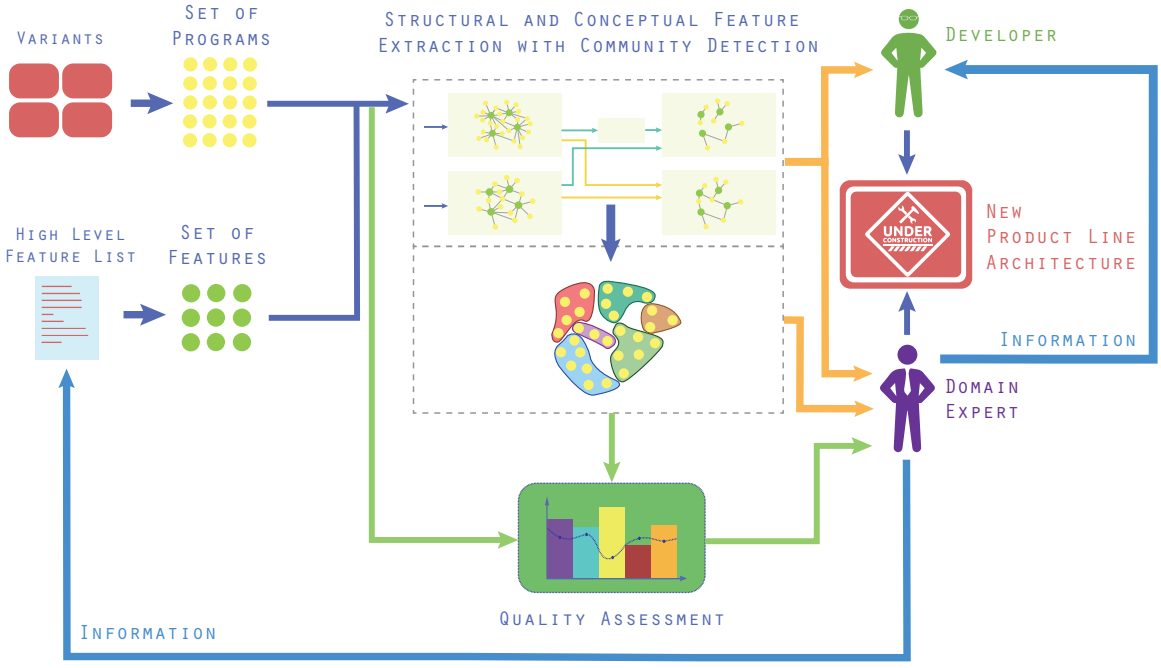

Figure 1: An illustration of feature extraction and analysis as part of product line adoption

Our feature extraction experiments produced various outputs suitable for different stages of the work. A pipeline of our process can be seen in Figure 1, which paints a high-level picture. These steps will be explained further.

Call graphs through static analysis highlight the structural connections within the software's programs and can produce an output fit for developers. Another method, information retrieval (IR) through latent semantic indexing (LSI), can also highlight program and feature connections by examining the similarities between the parts of the programs written in natural language and the features [19]. This is closer to the view of the domain experts, who typically do not delve deep into the programs but see the conceptual connections. The two extraction methods build on different sources of information. Their outputs could be combined to produce more strict filtering of programs that contain the most essential programs for each feature [20]. The sets of

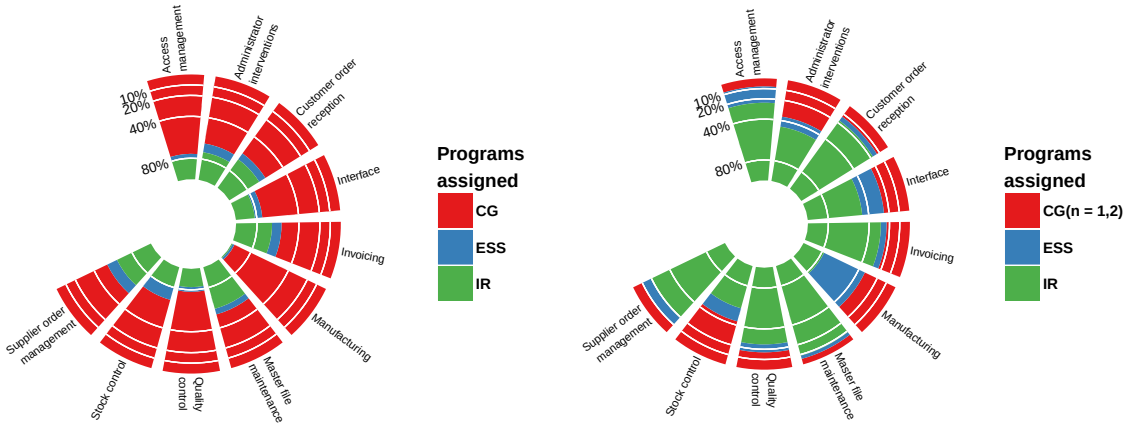

Figure 2: The size of our result sets for each feature with each technique. Results without filtering shown on the left, results with filtering on the right.

programs extracted from a variant with the two techniques are shown on the left side of Figure 2, where ESS represents the set that both techniques found. Each slice of the diagram represents a top-level feature (most general features), and its colors indicate the number of programs detected by each technique. Figure 3 displays a combined graph of the essential programs of the same variant for each feature. As visible in the images, they can be further filtered by determining a maximal number of features a program can connect before it is considered less specific and filtered out. It is apparent from the graph that features are much better separated, providing a suitable high-level glance at the background of features without a lot of technical details, ideal for top-level understanding.

Domain experts could further benefit from overseeing the connections between features. The call graph contains most of these connections and, while it is hard to comprehend, provides valid structural information. The call graph could benefit from methods commonly used in graph theory, namely community detection in our case. Communities could provide a good mapping of features and their interconnections within the system [10]. Our evaluation has shown that a community mapping can highlight connections between features that may not be apparent to the domain experts, while in general, still providing a picture close to their estimations.

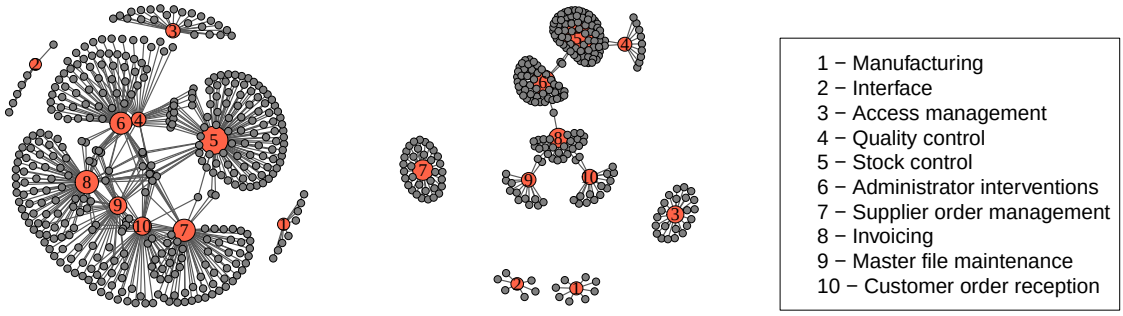

Figure 3: Graph visualization of the set of programs deemed most essential. Results shown on the left, results with filtering on the right

Many aspects of the features could potentially be useful to investigate during the feature extraction process and maintenance. While mainstream languages have the benefit of several well-established metrics, 4GL environments suffer from a lack of these. Our work contributed several new metrics that are adaptations of already established metrics but also some original metrics suitable for the abstraction level of features [6, 7]. Some of these new metrics are displayed in Figure 4 featuring four versions of the product line under composition through different stages of the work. In the figure, NP represents the number of programs, while HD depicts the Halstead Difficulty metric which is a measure of complexity.

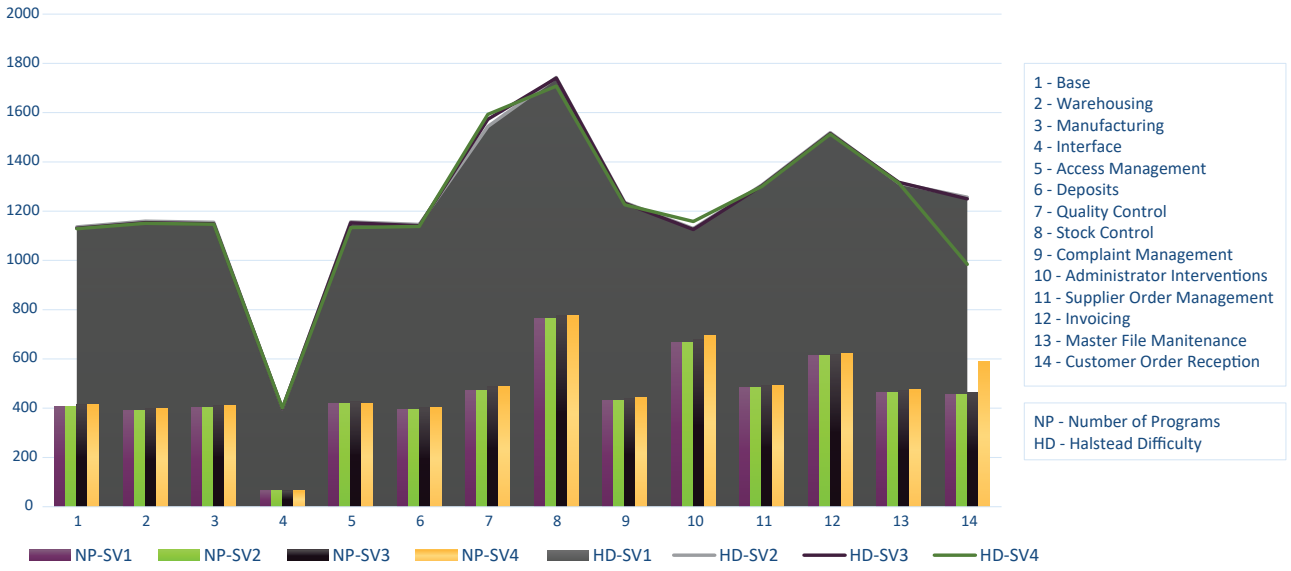

Figure 4: A comparison of the four versions of the system regarding their number of programs (NP) and their complexity (HD)

According to an evaluation by the domain experts and developers, our outputs hold valid feature extraction information. Our output combining structural and conceptual information contains the more relevant programs of a feature, with more than 70% valid matches with over 45% also being relevant overall. At least a third of these combined matches were relevant in every single case. The call graph communities were found to represent real feature coupling information that approximates domain expert knowledge well while relying only on structural information, thus can contribute to the decisions of a domain expert by providing another informed viewpoint. The methods and results of our work had been used during the work of our industrial partner, and the adoption process was concluded successfully.

The Author's Contributions

The author implemented the information retrieval feature-extraction solution and took part in the planning and coordination of the experiments, including the combination possibilities, new metrics, community detection, and evaluation. He took part in the combination effort between static analysis and information retrieval, both in implementation and the analysis of the results. The author implemented a basic feature-extractor with graphical interface based on information retrieval for the initial approach, and later performed the analysis of the variants, and planned the validation procedure. The publications related to this thesis point are:

- **András Kicsi**, Viktor Csuvik, László Vidács, Ferenc Horváth, Árpád Beszédes, Tibor Gyimóthy, and Ferenc Kocsis. Feature Analysis using Information Retrieval, Community Detection and Structural Analysis Methods in Product Line Adoption. Journal of Systems and Software, 155:70–90, sep 2019.

- **András Kicsi**, László Vidács, Arpád Beszédes, Ferenc Kocsis, and István Kovács. Information retrieval based feature analysis for product line adoption in 4gl systems. In Proceedins of the 17th International Conference on Computational Science and Its Applications – ICCSA 2017, pages 1–6. IEEE, 2017.

- **András Kicsi**, Viktor Csuvik, László Vidács, Árpád Beszédes, and Tibor Gyimóthy. Feature level complexity and coupling analysis in 4GL systems. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10964 LNCS, pages 438–453. Springer Verlag, may 2018.

- **András Kicsi**, László Vidács, Viktor Csuvik, Ferenc Horváth, Arpád Beszédes, and Ferenc Kocsis. Supporting product line adoption by combining syntactic and textual feature extraction. In International Conference on Software Reuse, ICSR 2018. Springer International Publishing, 2018.

- **András Kicsi** and Viktor Csuvik. Feature Level Metrics Based on Size and Similarity in Software Product Line Adoption. In 11th Conference of PhD Students in Computer Science (CSCS 2018), pages 25–28, 2018.

II Textual Methods in Aiding Test-to-Code Traceability

The second thesis point focuses on aiding test-to-code traceability through textual information. Test-to-Code traceability is the identification of each test case's focus, finding out which parts of the software they are meant to assess. While their research is less popular than, for instance, requirement traceability, there is substantial literature on the matter, and it is not an easy problem considering that larger systems can have tens of thousands of test cases. Proper test-to-code traceability can facilitate the localization of faults, aid test prioritization, and could even serve as a foundation for good automatic program repair. While the current state-of-the-art solutions tend to employ multiple approaches to the task, our focus was on the lexical techniques. Our research examined eight medium-sized open-source systems with more than 1.25 million lines of code. A high-level overview of our main process is visible in Figure 5, starting from the source files at the top left.

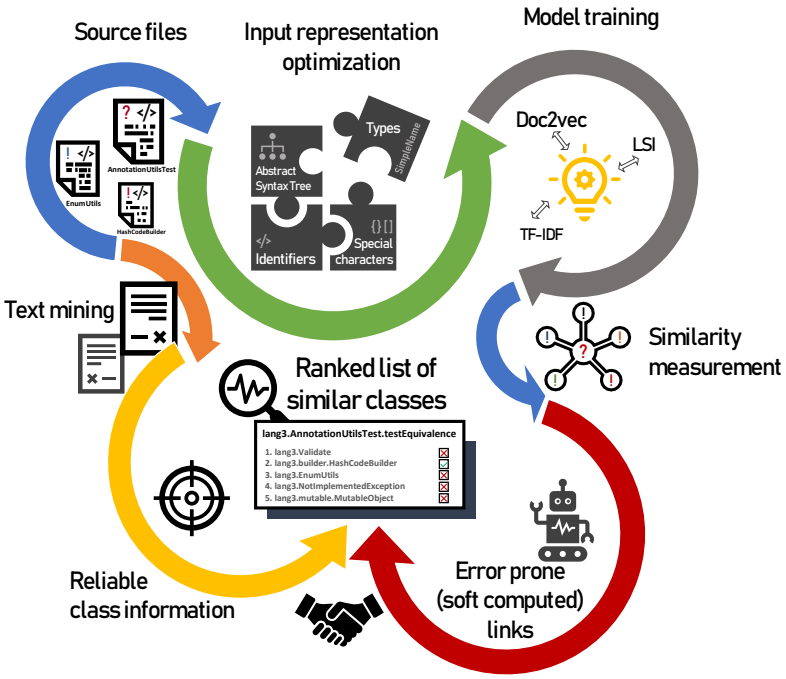

Figure 5: A high-level overview of the proposed process

Relying on naming conventions (NC) is an exceptionally precise way to identify the tested code elements, but this technique is highly dependent on developer habits and can be complicated in a variety of cases. A trivial example of naming conventions is visible in Figure 6. Our work provided an in-depth investigation about the automatic extraction of naming convention links, the possible combinations of various rules, and their perceived usage in the systems [8]. The results show that on the method level, naming conventions are complicated, and developers rarely strive to uphold perfect traceability. On the class level, they are used quite often, which can greatly contribute to class-level test-to-code traceability. Mirroring the package hierarchy of the production code is also popular, and even if this can only lead to general directions in the code, this can still be extremely useful as filtering information.

Since textual approaches are often used in state-of-the-art solutions, optimizing them could lead to better traceability overall. Thus, our experiments also searched for the best textual method. We introduced several new possible combinations and examined their usefulness in finding correct traceability links, measured both on a large set of traceability links automatically extracted according to naming conventions and a manual dataset on 220 test cases of four systems.

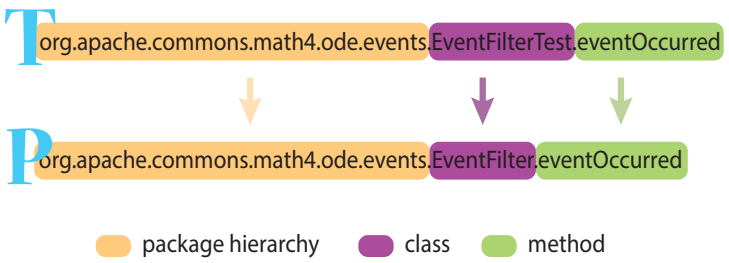

Figure 6: A trivial naming convention example from Commons Math

The precision results of the NC-based evaluation are visible in Table 1, with the best results highlighted. From standalone techniques, Doc2Vec seemed most reliable in the majority of the cases. Ensemble-50 represents the output of Doc2Vec filtered with the top 50 similarity results of both LSI and TF-IDF. It is visible that this combination was found to be actually worse than Doc2Vec by itself. Doc2Vec+CG filters with call information obtained via regular expressions. As it is visible, call information was found to contribute greatly to good results, reinforcing that a combination of techniques is indeed likely to produce even better results.

Table 1: Top-1 results featuring the different text-based models trained on various source code representations, evaluated using naming conventions. - highest value in a row - highest value in a column

Method	Representation	ArgoUML	C. Lang	C. Math	Gson	JFreeChart	Joda-Time	Mondrian	PMD
Doc2Vec	IDENT	19.63%	82.16%	50.00%	45.83%	49.22%	41.43%	66.42%	37.15%
	LEAF	18.43%	61.00%	33.01%	47.92%	25.10%	20.79%	65.33%	42.04%
	SIMPLE	24.77%	67.91%	33.78%	47.50%	30.69%	26.26%	65.33%	34.82%
	SRC	7.85%	31.32%	15.46%	16.67%	22.64%	22.30%	21.53%	15.92%
	TYPE	0.60%	4.36%	0.78%	2.92%	2.36%	5.18%	0.00%	0.00%
LSI	IDENT	32.93%	66.08%	19.42%	30.83%	33.29%	35.04%	22.99%	19.96%
	LEAF	14.80%	23.11%	3.63%	7.08%	9.63%	16.12%	11.31%	7.80%
	SIMPLE	15.71%	21.48%	3.47%	4.37%	13.15%	6.69%	4.38%	11.46%
	SRC	19.64%	54.64%	24.36%	14.17%	21.48%	28.20%	31.02%	22.29%
	TYPE	0.00%	0.48%	0.65%	4.58%	0.00%	0.50%	0.00%	0.00%
TF-IDF	IDENT	35.95%	73.62%	35.78%	35.00%	45.65%	48.71%	73.72%	24.63%
	LEAF	32.63%	70.94%	37.33%	38.33%	48.93%	47.77%	66.79%	23.99%
	SIMPLE	28.70%	69.49%	33.08%	30.00%	44.30%	47.77%	72.26%	23.57%
	SRC	27.79%	51.51%	28.68%	18.75%	25.19%	31.08%	50.73%	22.29%
	TYPE	0.00%	0.48%	0.65%	4.58%	0.00%	0.50%	0.00%	0.00%
Ensemble-50	IDENT	13.89%	48.00%	28.27%	27.92%	50.00%	30.22%	4.75%	33.97%
	LEAF	11.18%	31.61%	19.29%	33.75%	33.56%	24.89%	1.45%	35.03%
	SIMPLE	15.41%	28.54%	18.93%	38.75%	34.01%	22.73%	1.01%	25.48%
	SRC	6.04%	20.00%	11.02%	13.33%	23.24%	16.33%	1.46%	16.35%
	TYPE	0.00%	3.79%	0.12%	0.00%	1.11%	0.00%	0.00%	0.00%
Doc2Vec+CG	IDENT	45.01%	83.14%	61.04%	85.83%	62.82%	43.02%	68.61%	54.14%
	LEAF	42.29%	72.66%	45.65%	44.17%	56.48%	34.10%	73.72%	52.23%
	SIMPLE	41.69%	71.89%	51.41%	52.92%	58.06%	24.32%	74.09%	51.80%
	SRC	32.33%	54.48%	29.62%	35.42%	37.36%	23.38%	47.08%	36.31%
	TYPE	3.63%	17.61%	13.22%	42.08%	10.46%	16.04%	41.24%	15.92%

The precision results measured on TestRoutes, the set of 220 test cases, are featured in Table 2. The NC row represents the applicability of a naming convention that uses class and package name complete matches as a basis with an additional "test" string or package included.

As it is visible, Doc2Vec still seems to be the best standalone technique and call information is still a good supplement. Doc2Vec+CG+NC considers the naming convention as a basis, and provides the Doc2Vec+CG results in cases where the NC criteria are not met. As visible, this method performed best evaluated on manual data.

Table 2: Top-1 and top-5 results featuring the different text-based models and the applicability of NC on each project. Models were trained on 5 different source code representations. - highest value in a row - highest value in a column

Method	Representation	Top-1				Top-5			
		C. Lang	Gson	JFreeChart	Joda-Time	C. Lang	Gson	JFreeChart	Joda-Time
NC	-	76.00%	26.00%	0.00%	0.00%	76.00%	26.00%	0.00%	0.00%
Doc2Vec	IDENT	58.00%	15.69%	15.49%	32.00%	62.00%	25.49%	15.49%	48.00%
	LEAF	30.00%	13.73%	11.27%	20.00%	52.00%	21.57%	15.49%	52.00%
	SIMPLE	15.69%	17.65%	14.08%	16.00%	48.00%	21.57%	16.90%	52.00%
	SRC	16.00%	9.80%	12.68%	32.00%	42.00%	29.41%	30.99%	54.00%
	TYPE	4.00%	1.96%	11.27%	4.00%	22.00%	3.92%	11.27%	10.00%
LSI	IDENT	34.00%	17.65%	4.23%	10.00%	68.00%	5.64%	5.63%	44.00%
	LEAF	12.00%	7.84%	4.23%	2.00%	34.00%	23.53%	5.63%	28.00%
	SIMPLE	4.00%	5.88%	4.23%	2.00%	30.00%	23.53%	5.63%	24.00%
	SRC	34.00%	17.65%	12.68%	20.00%	70.00%	37.25%	23.94%	58.00%
	TYPE	4.00%	0.00%	0.00%	0.00%	8.00%	64.71%	0.00%	14.00%
TF-IDF	IDENT	30.00%	19.61%	4.23%	46.00%	76.00%	31.37%	5.63%	70.00%
	LEAF	30.00%	19.61%	4.23%	44.00%	76.00%	33.33%	5.63%	70.00%
	SIMPLE	28.00%	21.57%	4.23%	44.00%	72.00%	33.33%	5.63%	72.00%
	SRC	38.00%	19.61%	23.94%	12.00%	78.00%	43.14%	25.35%	68.00%
	TYPE	4.00%	0.00%	0.00%	0.00%	8.00%	64.71%	0.00%	14.00%
Ensemble-50	IDENT	44.00%	13.73%	4.23%	6.00%	52.00%	23.53%	4.23%	10.00%
	LEAF	13.73%	13.73%	4.23%	10.00%	38.00%	19.61%	4.23%	14.00%
	SIMPLE	14.00%	15.69%	4.23%	2.00%	40.00%	19.61%	4.23%	8.00%
	SRC	7.84%	11.76%	11.27%	12.00%	36.00%	17.65%	28.17%	22.00%
	TYPE	2.00%	1.96%	0.00%	2.00%	8.00%	1.96%	0.00%	2.00%
Doc2Vec+CG	IDENT	58.00%	64.71%	16.90%	24.00%	76.00%	80.39%	23.94%	64.00%
	LEAF	54.00%	54.90%	18.31%	20.00%	72.00%	78.43%	33.80%	66.00%
	SIMPLE	50.00%	56.86%	25.35%	26.00%	76.00%	78.43%	45.07%	64.00%
	SRC	50.00%	56.86%	36.62%	32.00%	78.00%	82.35%	66.19%	74.00%
	TYPE	42.00%	47.05%	11.27%	6.00%	62.00%	74.51%	28.17%	24.00%
Doc2Vec+CG+NC	IDENT	76.00%	64.71%	16.90%	24.00%	86.00%	72.55%	23.94%	64.00%
	LEAF	78.00%	64.71%	18.31%	20.00%	84.00%	70.59%	33.80%	66.00%
	SIMPLE	78.00%	66.71%	25.35%	26.00%	84.00%	76.47%	45.07%	64.00%
	SRC	80.00%	66.67%	36.62%	32.00%	88.00%	78.43%	66.19%	74.00%
	TYPE	74.00%	64.71%	11.27%	6.00%	78.00%	76.47%	28.17%	24.00%

Our comparison of five source code representations, visible in both tables as different rows, provided less conclusive results. The identifier-centric (IDENT) representation that utilizes abstract syntax trees came out on top in the overwhelming majority of the cases during the NC-based evaluation, but the text-centric (SRC) representation proved more precise when compared to the limited amount of manual data.

Out of the main methods of latent semantic indexing, TF-IDF and Doc2Vec, Doc2Vec was found to be most precise both as a singular technique and in combination with call information or naming conventions. The combination of these three techniques could also warrant some attention, but our filtering solution based on a consensus of all three of them provided marginally worse results than Doc2Vec in itself.

Our findings show that the combination of naming conventions and Doc2Vec could lead to a good alloy of the precision of naming conventions and the versatility of other textual techniques.

Our work also contributed several previous results leading to this evaluation [2, 1, 17], the TestRoutes manual dataset [18], and traceability extraction experiments on Stack Overflow code snippets featuring LSI an Doc2Vec, in which Doc2Vec seemed to perform better [13].

The Author's Contributions

The author implemented a Latent Semantic Indexing based solution for recovering traceability links, the evaluation code relying on naming conventions and also manual data. He also implemented recovery techniques based on various naming conventions and conducted experiments with them. The author planned and coordinated the manual annotation of the TestRoutes dataset, and also the Stack Overflow experiments. He also took part in the evaluation and explanation of various other results and the planning of all of the published experiments. The publications related to this thesis point are:

- **András Kicsi**, Viktor Csuvik, and László Vidács. Large Scale Evaluation of NLP-based Test-to-Code Traceability Approaches. IEEE Access, 2021.
- **András Kicsi**, László Tóth, and László Vidács. Exploring the benefits of utilizing conceptual information in test-to-code traceability. Proceedings of the 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, pages 8–14, 2018.
- Viktor Csuvik, **András Kicsi**, and László Vidács. Source code level word embeddings in aiding semantic test-to-code traceability. In 10th International Workshop at the 41st International Conference on Software Engineering (ICSE) – SST 2019. IEEE, 2019.
- Viktor Csuvik, **András Kicsi**, and László Vidács. Evaluation of Textual Similarity Techniques in Code Level Traceability. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 11622 LNCS, pages 529–543. Springer Verlag, 2019.
- **András Kicsi**, Márk Rákóczi, and László Vidács. Exploration and mining of source code level traceability links on stack overflow. In ICSOFT 2019 - Proceedings of the 14th International Conference on Software Technologies, pages 339–346, 2019.
- **András Kicsi**, László Vidács, and Tibor Gyimothy. Testroutes: A manually curated method level dataset for test-to-code traceability. In Proceedings of the 17th International Conference on Mining Software Repositories, MSR 2020, pages 593–597. IEEE, IEEE, jun 2020.

III Machine Understanding of Radiologic Reports

The third thesis point focuses on the machine understanding of Hungarian radiologic spinal reports. Radiologic examinations produce image data, but their main output that incorporates the expertise of radiologists is still manifested in textual form. Reports and opinions written by the radiologists contain significant information which could be utilized in various services such as quality assurance and automatic report generation. The typical process of the composition of radiologic reports can be seen in Figure 7.

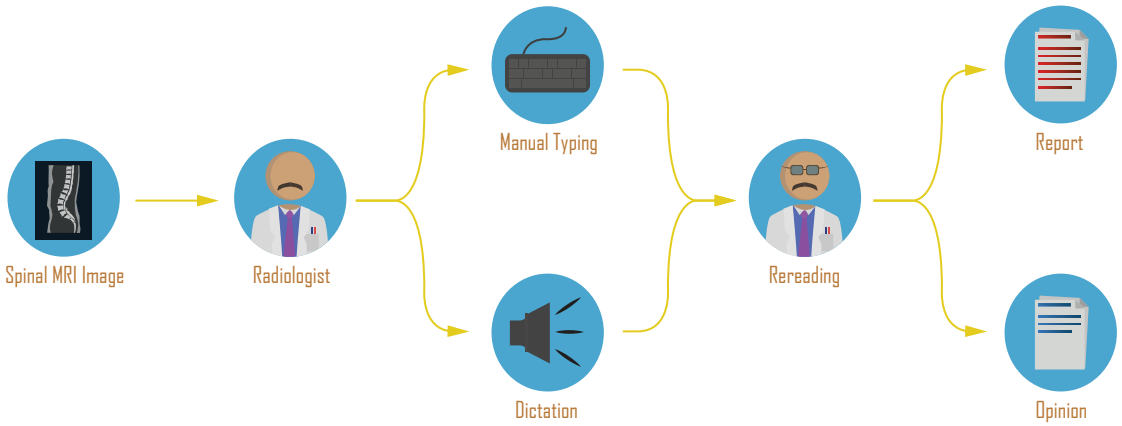

Figure 7: The workflow of a radiologic examination

The thesis point describes a process for the automatic understanding of radiologic reports of the spinal region. This process involves multiple levels, which are elaborated below.

Radiologic reports tend to contain a high number of misspellings. While these are relatively easy to overlook by the human eye, they can inhibit the work of machine learning classifiers, as well as rule-based methods. Correcting these automatically is a hard task not only because of the morphologic richness of the Hungarian language but also because radiologic reports tend to use Latin words according to the rules of other Hungarian terms, attaching Hungarian suffixes. Our work introduced a method based on the HunSpell spelling correction tool as well as its considerable extension with a specialized dictionary and new prioritization rules to accommodate the specific use of the language in the reports [14]. Our method is shown to improve the results of both machine learning based classification (0.35 F-score improvement) and ontology-based identification (more than 20% of the unknown terms correctly identified).

The corrected natural language text goes through automatic labelling via a BiLSTM-CRF [12], or in the more current version, a BERT classifier. It distinguishes anatomical locations, disorders and properties according to an annotation performed by radiologists on 487 real reports. The classification produced an F1-score value of 96.85, determining locations, disorders and properties with high accuracy. An English language illustration of this classification is visible on the left side of Figure 8. The consistency of the annotation was refined by the original radiologist a statistical helper tool [11] and was consolidated with another radiologist's annotation.

Our identification relies on a simple ontology built for the task. The ontology lists identifiers for various locations with a simple hierarchy, providing the vertebra level of anatomical locations if they are applicable (for example, the disc at the L5 level is denoted as L5_D). The identifiers of disorders also carry additional information as they are grouped in harmful pathologies, describing a normal or intact state and aspects that are needed to specify another accompanying disorder but are meaningless without it. Properties can be less easily generalized by such simple rules and they are not identified at the current time.

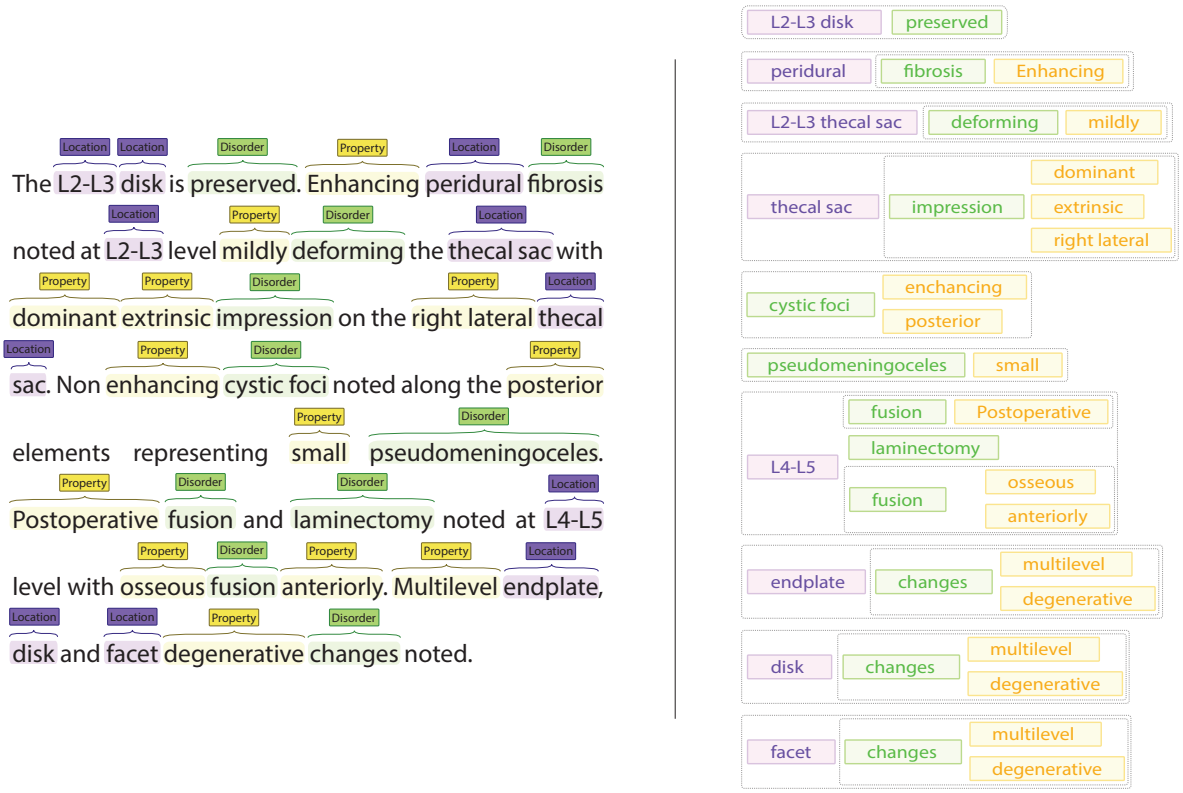

Figure 8: Left: An English language illustration of our annotation system. Right: An illustration of our structured visualization of this text.

Negations are determined through linguistic analysis via the Magyarlánc [24] analyzer. Constituent parsing is used to link negating words to specific disorders in the text, detected by our classification.

Our process also maps the elements according to their semantic connections. Disorders are usually connected with one or more locations, and properties usually belong to one or more disorders. Pairs of locations or disorders are also matched. Our system uses a rule-based method to determine connections, also heavily relying on the clauses of the sentences, extracted via constituents [15].

The output of our process is visualized in an easy to comprehend tree-structure that showcases the detected elements and their connections [16]. This output is displayed on the right side of Figure 8. The whole process is illustrated in Figure 9.

A comparison of our BiLSTM-CRF and BERT-based classification models is visible in Table 3, showing that the huBERT [23] provides significantly better results measured on 20% of the 487 annotated reports.

Table 3: A comparison of our previous BiLSTM-CRF model and our new BERT-based entity classification

	Accuracy	**F-score**
BertForTokenClassification	96.98	96.85
BiLSTM-CRF	95.10	95.09

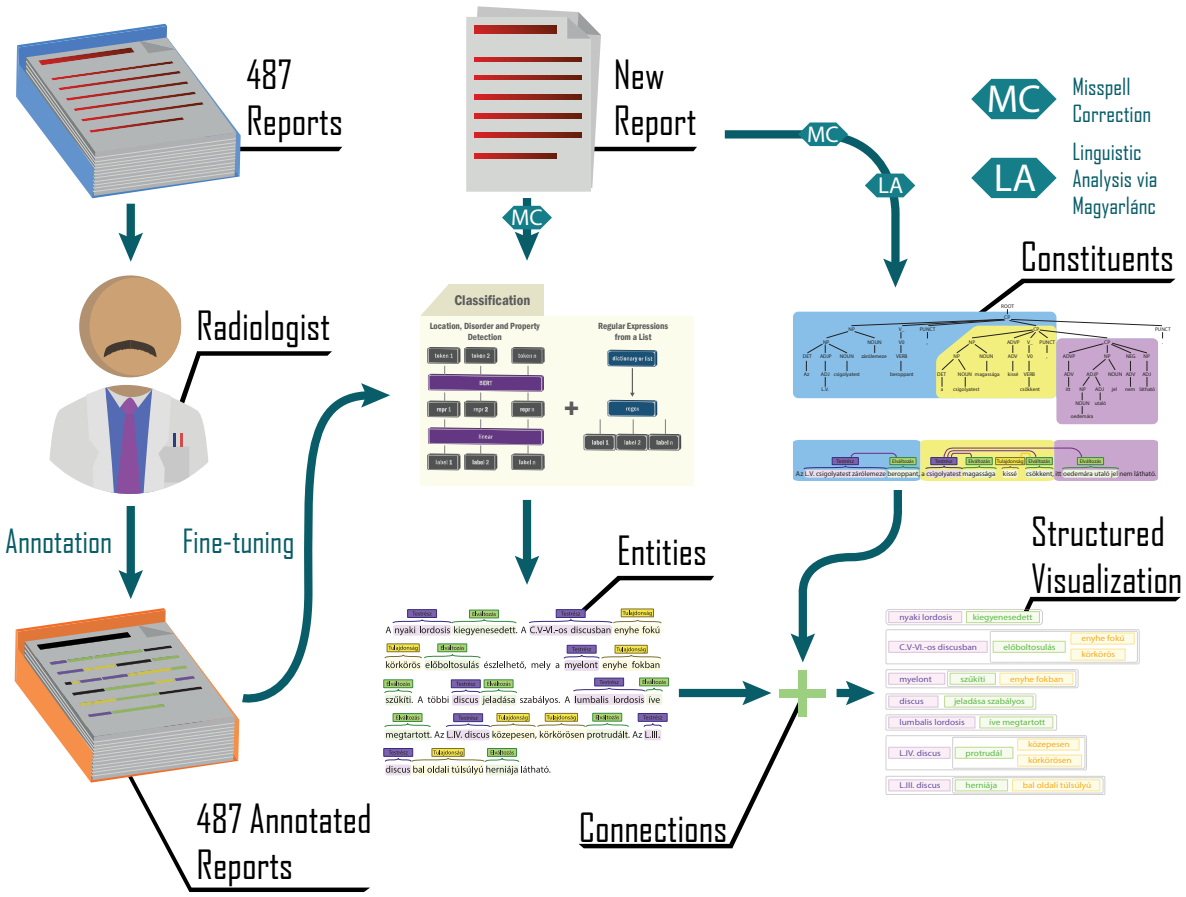

Figure 9: Our proposed method for automatic understanding and visualization

In a functional evaluation, three radiologists evaluated our machine understanding concerning classifications, the determination of connections, and the identification of locations and disorders. They determined the maximum achievable points according to functional evaluation guidelines and also did the scoring of each sentence, treating the three levels of the process separately. As shown in Table 4, the radiologists (R1, R2 and R3) found that our process provided correct results in 95.97% to 97.74% of the cases.

Table 4: The results of the understanding scores according the three radiologists

Task	R1	R2	R3	Average
Classification	97.61%	97.23%	98.47%	97.75%
Connection	96.22%	98.29%	91.37%	95.32%
Identification	98.48%	97.92%	96.75%	97.72%
Overall	97.57%	97.74%	95.97%	97.12%

While this is still ongoing research, some applications have already arisen. An automatic understanding of the reports could facilitate the generation of training data for automatic analyzer tools aiming to detect disorders on MRI images. As part of a current project, our solution is used for such a task. Another evident use of our work is the facilitation of patient comprehension. Interactive electronic reports could provide much more information for patients and lead to better healthcare service, and our project also includes this endeavor.

The Author's Contributions

The author laid the groundwork and coordinated the manual annotations, and took a big part in the later refinement of the data. He took part in the planning of all aspects of the machine understanding method and coordinated its implementation. The author planned the functional evaluation and its guidelines. He took a big role in the evaluation and explanation of the results and their implications. The publications related to this thesis point are:

- **András Kicsi**, Klaudia Szabó Ledenyi, Péter Pusztai, and László Vidács. Automatic classification and entity relation detection in hungarian spinal MRI reports. In 3rd ICSE Workshop on Software Engineering for Healthcare, 2021.
- **András Kicsi**, Péter Pusztai, Klaudia Szabó Ledenyi, Endre Szabó, Gábor Berend, Veronika Vincze, and László Vidács. Információkinyerés magyar nyelvű gerinc mr leletekből. In XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019), page 177–186, Szeged, 2019. (In Hungarian)
- **András Kicsi**, Klaudia Szabó Ledenyi, Péter Pusztai, Péter Németh, and László Vidács. Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 15–28, Szeged, 2020. (In Hungarian)
- **András Kicsi**, Klaudia Szabó Ledenyi, Péter Németh, Péter Pusztai, László Vidács, and Tibor Gyimóthy. Elírások automatikus detektálása és javítása radiológiai leletek szövegében. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 191–204, Szeged, 2020. (In Hungarian)
- **András Kicsi**, Péter Pusztai, Endre Szabó, and László Vidács. Szaknyelvi annotációk javításának statisztikai alapú támogatása. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 115–128, Szeged, 2020. (In Hungarian)

Summary

The thesis discusses three main topics. The first part provides insight into a project conducted with an industrial partner whose 4GL system has undergone software product line adoption. The thesis elaborates on how our work contributed to the project and showcases our solutions on the analysis of the variants of their system. While product line adoption is a well-researched topic, 4GL solutions are still rare, with little literature. Our work contributed with novel solutions on feature extraction based on static analysis and information retrieval and investigated their combinations. This part of the thesis also showcases several new metrics suitable for the analysis of the features of 4GL systems and a community-based method that groups parts of the software according to their mutual connections.

The second part describes the importance of textual methods in test-to-code traceability and how improving them could lead to the improvement of the whole process. The thesis analyzes some of the commonly used techniques, such as naming conventions and LSI, and provides a new alternative in Doc2Vec and several combinations of these, including various source code representations. It has been established that methods for this field work best in combination, and the textual aspect is likely to remain an important part of future solutions.

The third part describes how radiologic reports are made in today's medicinal practice and how their automatic understanding could contribute to better healthcare and future work in the field of machine learning. It establishes our work in the classification of various entities via BiLSTM-CRF and BERT models in the non-structured text of the reports and their connection through linguistic analysis. These elements are connected to a simple ontology as means of proper identification, and a tree-structured representation is constructed of them. Misspellings can hinder this identification, therefore the thesis point also described a solution for their automatic correction.

The future still holds many more interesting directions for these topics. Software reuse is flourishing in the industrial setting and is unlikely to ever lose its importance as it is often more cost-effective to build upon earlier achievements. Even as new solutions for test-to-code traceability tend to produce great results, there is still ample room for improvement. One such improvement can be to upgrade parts of our current methods as our research has proposed. Our work on radiologic reports is merely the beginning of an effort aiming to provide smarter, more optimised, and less resource-intensive healthcare. The automatic understanding can be extended to different parts of the body or different languages, and the vast amount of knowledge already in the reports can be extracted to enable new methods for more ambitious goals.

Table 5 provides an overview of the author's publications related to each thesis point.

№	[19]	[9]	[7]	[20]	[10]	[17]	[2]	[1]	[13]	[18]	[8]	[12]	[15]	[14]	[11]	[16]
I.	◆	◆	◆	◆	◆											
II.						◆	◆	◆	◆	◆	◆					
III.												◆	◆	◆	◆	◆

Table 5: Thesis points and supporting publications

Acknowledgments

There are so many who helped me along the journey of my studies that I couldn't possibly name them all. My utmost gratitude goes out to my supervisor, László Vidács, who, in my opinion, is the best supervisor I could have ever wished for, and not just for my doctoral studies, but also my work. I will be forever grateful for all the guidance and care he provided through the years. I do not consider myself easily inspired by outside sources, and yet, to this day, he can always manage this. I am exceptionally thankful for my amazing co-authors and my dear colleagues who greatly contributed to the success of our research. To name a few of them, I would like to thank Viktor Csuvik, Klaudia Szabó Ledenyi, Péter Pusztai, and Ferenc Horváth without whom I would have much less to write about now, as their devoted work was indispensable for my own. I would also like to thank Tibor Gyimóthy for providing me with an offer for doctoral studies at the Department of Software Engineering, and many interesting research opportunities ever since. It was a privilege to conduct my studies in such company.[1]

I would also like to thank my family, who have given me a brain and a heart to face the world, and have continuously supported me through it. I am extremely grateful for my love and my wonderful friends as well, who truly give merit to my life.

András Kicsi, 2022

[1]My work was also supported by the ÚNKP-21-4-1 and ÚNKP-22-4-1 New National Excellence Program of the Ministry for Innovation and Technology from the Source of the National Research, Development and Innovation Fund.

References

[1] Viktor Csuvik, András Kicsi, and László Vidács. Evaluation of Textual Similarity Techniques in Code Level Traceability. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11622 LNCS, pages 529–543. Springer Verlag, 2019.

[2] Viktor Csuvik, András Kicsi, and László Vidács. Source code level word embeddings in aiding semantic test-to-code traceability. In *10th International Workshop at the 41st International Conference on Software Engineering (ICSE) – SST 2019*. IEEE, 2019.

[3] S C Deerwester, S T Dumais, T K Landauer, G W Furnas, and R A Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[4] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, pages 4171–4186, 2019.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[6] A. Kicsi, V. Csuvik, L. Vidács, Á. Beszédes, and T. Gyimóthy. Feature level complexity and coupling analysis in 4GL systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10964 LNCS, pages 438–453. Springer, Cham, may 2018.

[7] András Kicsi and Viktor Csuvik. Feature Level Metrics Based on Size and Similarity in Software Product Line Adoption. In *11th Conference of PhD Students in Computer Science (CSCS 2018)*, pages 25–28, 2018.

[8] András Kicsi, Viktor Csuvik, and László Vidács. Large Scale Evaluation of NLP-based Test-to-Code Traceability Approaches. *IEEE Access*, 2021.

[9] András Kicsi, Viktor Csuvik, László Vidács, Árpád Beszédes, and Tibor Gyimóthy. Feature level complexity and coupling analysis in 4GL systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10964 LNCS, pages 438–453. Springer Verlag, may 2018.

[10] András Kicsi, Viktor Csuvik, László Vidács, Ferenc Horváth, Árpád Beszédes, Tibor Gyimóthy, and Ferenc Kocsis. Feature Analysis using Information Retrieval, Community Detection and Structural Analysis Methods in Product Line Adoption. *Journal of Systems and Software*, 155:70–90, sep 2019.

[11] András Kicsi, Péter Pusztai, Endre Szabó, and László Vidács. Szaknyelvi annotációk javításának statisztikai alapú támogatása. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020)*, page 115–128, Szeged, 2020.

[12] András Kicsi, Péter Pusztai, Klaudia Szabó Ledenyi, Endre Szabó, Gábor Berend, Veronika Vincze, and László Vidács. Információkinyerés magyar nyelvű gerinc mr leletekből. In *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)*, page 177–186, Szeged, 2019.

[13] András Kicsi, Márk Rákóczi, and László Vidács. Exploration and mining of source code level traceability links on stack overflow. In *ICSOFT 2019 - Proceedings of the 14th International Conference on Software Technologies*, pages 339–346, 2019.

[14] András Kicsi, Klaudia Szabó Ledenyi, Péter Németh, Péter Pusztai, László Vidács, and Tibor Gyimóthy. Elírások automatikus detektálása és javítása radiológiai leletek szövegében. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020)*, page 191–204, Szeged, 2020.

[15] András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, Péter Németh, and László Vidács. Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020)*, page 15–28, Szeged, 2020.

[16] András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, and László Vidács. Automatic classification and entity relation detection in hungarian spinal mri reports. In *3rd ICSE Workshop on Software Engineering for Healthcare*, 2021.

[17] András Kicsi, László Tóth, and László Vidács. Exploring the benefits of utilizing conceptual information in test-to-code traceability. *Proceedings of the 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering*, pages 8–14, 2018.

[18] András Kicsi, László Vidács, and Tibor Gyimothy. Testroutes: A manually curated method level dataset for test-to-code traceability. In *Proceedings of the 17th International Conference on Mining Software Repositories, MSR 2020*, pages 593–597. IEEE, IEEE, jun 2020.

[19] András Kicsi, László Vidács, Arpád Beszédes, Ferenc Kocsis, and István Kovács. Information retrieval based feature analysis for product line adoption in 4gl systems. In *Proceedins of the 17th International Conference on Computational Science and Its Applications – ICCSA 2017*, pages 1–6. IEEE, 2017.

[20] András Kicsi, László Vidács, Viktor Csuvik, Ferenc Horváth, Arpád Beszédes, and Ferenc Kocsis. Supporting product line adoption by combining syntactic and textual feature extraction. In *International Conference on Software Reuse, ICSR 2018*. Springer International Publishing, 2018.

[21] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001.

[22] Tomas Mikolov, Ilya Sutskever, Kan Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2:3111–3119, dec 2013.

[23] Dávid Márk Nemeskey. Introducing huBERT. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 3–14, 2019.

[24] János Zsibrita, Veronika Vincze, and Richárd Farkas. Magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In *International Conference Recent Advances in Natural Language Processing, RANLP*, pages 763–771, 2013.

Összefoglaló

Az emberiség legfőbb információátviteli módja az írott vagy beszélt nyelv. Az értekezés a természetesnyelvű szövegben rejlő információ kinyerésével foglalkozik. Két tézispont szorosan kapcsolódik a szoftverfejlesztés témaköréhez, ahol a programkódban szintén óriási mennyiségű kiaknázható szemantikus információ van, míg a harmadik tézispont radiológiai leletek automatizált értelmezésével foglalkozik.

Számos variáns felett a szoftver-termékcsaládok bevezetése felettébb erőforrás-igényes feladat lehet, ennek egyik alapfeladata a funkcionalitások (feature-ök) kinyerése. Munkánk során egy negyedik generációs (4GL) nyelven írt gyógyszeripari logisztikai rendszer 19 variánsa feletti feature-kinyerésben vettünk részt. A feature-kinyerési kísérleteink többféle különböző kimenetet produkáltak, amelyek statikus elemzéssel kinyert hívási gráfokkal, szöveges információkinyeréssel, és ezek kombinációjával igyekeztek támogatni a fejlesztők és területi szakértők munkáját egyaránt. A hívási gráf alapján továbbá egy közösségi algoritmusokra alapuló elemzést is végeztünk. Számos új 4GL feature metrikát is bevezettünk, amelyek megkönnyíthetik az elemzést.

A teszt-kód nyomonkövethetőség annak azonosítása, hogy a tesztek a szoftver mely kódrészletét igyekeznek tesztelni. Bár a jelenlegi legkorszerűbb megoldások általában módszerek kombinációját használják, kutatási célunk a szöveges technikák vizsgálata volt. Nyolc közepes méretű, nyílt forrású rendszeren végeztünk méréseket, amelyek összesen több mint 1,25 millió kódsort öleltek fel. A névkonvenciók (NC) terén mélyreható elemzést végeztünk automatikus kinyerésüknek, más módszerekkel való kombinációjuknak, és használati szokásaiknak területein. Kísérleteinkkel a legjobb szöveges hasonlósági módszert is igyekeztünk felkutatni. Bemutattunk több új lehetséges kombinációt a meglévő lexikális módszerek fölött, és megvizsgáltuk ezek pontosságát mind egy nagy, névkonvenciók alapján automatizáltan kinyert, mind egy 220 tesztesetből álló kézi adathalmazon. Kísérleteink alapján a korábbiakban ismertnél jelentősen jobb eredmények érhetők el szöveges módszerekkel ezen a területen.

A radiológiai vizsgálatokból képi adat készül, de fő kimenetük mégis a szöveges lelet. Munkánk magyar nyelvű radiológiai gerincleletek automatizált értelmezésére mutat be egy módszert. A bemenetként kapott szöveg elírásait automatizáltan javítottuk, majd a szövegben különböző entitásokat (testrészek, elváltozások és tulajdonságok) klasszifikáltunk gépi tanulás segítségével. Ezek összekapcsolását nyelvi elemzéssel és egyéni szabályokkal végeztük, és a tagadásokat is felismertük a szövegben. A talált entitásokat egy egyszerű ontológia alapján azonosítottuk. Módszerünk kimenete egy áttekinthető fa-struktúra a lelet értelmezésével.

A szöveg mindenhol ott van életünkben, megfelelő feldolgozása pedig hatalmas lehetőségekkel kecsegtet. Mindhárom ismertetett irány tele van még új lehetőségekkel.

Nyilatkozat

Kicsi András *Utilization of Underlying Semantic Information in Textual Data* című PhD disszertációjában a következő eredményekben **Kicsi András** hozzájárulása volt a meghatározó:

Az **első tézisponthoz** (*Feature-Extraction of Magic Applications*) tartozó eredmények:

1. Szöveges hasonlóság mérése és konfigurációja Latent Semantic Indexing technikával a feature-fa és a Magic programok szövege között, a méréshez a programkód implementációja. A szemantikai hasonlóság implementációja, mint ajánlórendszer. A végső módszer egyik felét képező szöveghasonlósági eredmények szolgáltatása.
    - András Kicsi, László Vidács, Árpád Beszédes, Ferenc Kocsis, and István Kovács. Information retrieval based feature analysis for product line adoption in 4gl systems. In Proceedins of the 17th International Conference on Computational Science and Its Applications – ICCSA 2017, pages 1–6. IEEE, 2017.
    - András Kicsi, Viktor Csuvik, László Vidács, Árpád Beszédes, and Tibor Gyimóthy. Feature level complexity and coupling analysis in 4GL systems. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10964 LNCS, pages 438–453. Springer Verlag, may 2018.
    - András Kicsi, Viktor Csuvik, László Vidács, Ferenc Horváth, Árpád Beszédes, Tibor Gyimóthy, and Ferenc Kocsis. Feature Analysis using Information Retrieval, Community Detection and Structural Analysis Methods in Product Line Adoption. Journal of Systems and Software, 155:70–90, sep 2019.
    - András Kicsi, László Vidács, Viktor Csuvik, Ferenc Horváth, Arpád Beszédes, and Ferenc Kocsis. Supporting product line adoption by combining syntactic and textual feature extraction. In International Conference on Software Reuse, ICSR 2018. Springer International Publishing, 2018.
    - András Kicsi and Viktor Csuvik. Feature Level Metrics Based on Size and Similarity in Software Product Line Adoption. In 11th Conference of PhD Students in Computer Science (CSCS 2018), pages 25–28, 2018.
2. Feature-kinyerés prototípus implementációja, kezdetleges grafikus felülettel, futtatási opciókkal, és az eredmények értékelésével.
    - András Kicsi, László Vidács, Arpád Beszédes, Ferenc Kocsis, and István Kovács. Information retrieval based feature analysis for product line adoption in 4gl systems. In Proceedins of the 17th International Conference on Computational Science and Its Applications – ICCSA 2017, pages 1–6. IEEE, 2017.
3. Az új és adaptált metrikák eredményességének elbírálása, eredmények áttekintése és magyarázata, következetesek levonása a publikációhoz.
    - András Kicsi, Viktor Csuvik, László Vidács, Árpád Beszédes, and Tibor Gyimóthy. Feature level complexity and coupling analysis in 4GL systems. In Lecture Notes

in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10964 LNCS, pages 438–453. Springer Verlag, may 2018.

4. Szöveges és szerkezeti adatok együttes felhasználásának tervezése, az ezekkel kapcsolatos munka során az eredmények áttekintése és értelmezése, a megfelelő konfigurációk kiválasztása.
    - András Kicsi, László Vidács, Viktor Csuvik, Ferenc Horváth, Arpád Beszédes, and Ferenc Kocsis. Supporting product line adoption by combining syntactic and textual feature extraction. In International Conference on Software Reuse, ICSR 2018. Springer International Publishing, 2018.
5. Közösségi algoritmusok használatának tervezése, eredményeik értelmezése, szükséges konfigurációk megállapítása. A területi szakértők általi validáció kézi kiértékelésének előkészítése, a kiértékelés eredményeinek értelmezése, konklúziók levonása. A variánsok felett mért mérőszámok összehasonlítása, eredményeik értelmezése, vizualizációja.
    - András Kicsi, Viktor Csuvik, László Vidács, Ferenc Horváth, Árpád Beszédes, Tibor Gyimóthy, and Ferenc Kocsis. Feature Analysis using Information Retrieval, Community Detection and Structural Analysis Methods in Product Line Adoption. Journal of Systems and Software, 155:70–90, sep 2019.
6. 4GL Feature-hasonlósági metrikák tervezése és implementációja szemantikai hasonlóság és szerkesztési távolságok alapján. Hasonlósági metrikák eredményének értelmezése és megjelenítése.
    - András Kicsi and Viktor Csuvik. Feature Level Metrics Based on Size and Similarity in Software Product Line Adoption. In 11th Conference of PhD Students in Computer Science (CSCS 2018), pages 25–28, 2018.

A **második tézisponthoz** (*Test-to-Code Traceability*) tartozó eredmények:

1. Szemantikai hasonlóság mérésének implementációja Latent Semantic Indexing módszerrel a tesztek és kódosztályok, valamint kód-metódusok között.
    - András Kicsi, László Tóth, and László Vidács. Exploring the benefits of utilizing conceptual information in test-to-code traceability. Proceedings of the 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, pages 8–14, 2018.
    - Viktor Csuvik, András Kicsi, and László Vidács. Source code level word embeddings in aiding semantic test-to-code traceability. In 10th International Workshop at the 41st International Conference on Software Engineering (ICSE) – SST 2019. IEEE, 2019.
2. Névkonvenció-alapú kiértékelés megtervezése. Nyomonkövethetőségi eredmények kiértékelése, kiértékelő modul implementációja a névkonvenciók felhasználásával. Különböző kiértékelési módok kidolgozása és implementációja.
    - András Kicsi, László Tóth, and László Vidács. Exploring the benefits of utilizing conceptual information in test-to-code traceability. Proceedings of the 6th

International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, pages 8–14, 2018.

- Viktor Csuvik, András Kicsi, and László Vidács. Source code level word embeddings in aiding semantic test-to-code traceability. In 10th International Workshop at the 41st International Conference on Software Engineering (ICSE) – SST 2019. IEEE, 2019.
- Viktor Csuvik, András Kicsi, and László Vidács. Evaluation of Textual Similarity Techniques in Code Level Traceability. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 11622 LNCS, pages 529–543. Springer Verlag, 2019.
- András Kicsi, Viktor Csuvik, and László Vidács. Large Scale Evaluation of NLP-based Test-to-Code Traceability Approaches. IEEE Access, 2021.

3. LSI eredmények mérése, kombinációja név-egyezéssel. Ajánlórendszerként való felhasználás, ennek kiértékelése, magyarázata.
    - András Kicsi, László Tóth, and László Vidács. Exploring the benefits of utilizing conceptual information in test-to-code traceability. Proceedings of the 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, pages 8–14, 2018.
4. A különböző technikák eredményeinek összehasonlítása, eredmények értelmezése, magyarázata.
    - Viktor Csuvik, András Kicsi, and László Vidács. Evaluation of Textual Similarity Techniques in Code Level Traceability. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 11622 LNCS, pages 529–543. Springer Verlag, 2019.
5. StackOverflow teszt-kód nyomonkövethetőségi kísérletek megtervezése, felügyelete, koordinációja. Doc2Vec eredmények értelmezése, kiértékelő összehasonlítás elvégzése, következtetések levonása.
    - András Kicsi, Márk Rákóczi, and László Vidács. Exploration and mining of source code level traceability links on stack overflow. In ICSOFT 2019 - Proceedings of the 14th International Conference on Software Technologies, pages 339–346, 2019.
6. A publikált adathalmazhoz a kézi annotátor felkészítése a munkára, adatok szolgáltatása, és az adathalmazt képező tesztek kiválasztása. Folyamatos kapcsolattartás az annotátorral, munkájának felügyelete. Az annotáció eredményének ellenőrzése, kétséges elemek cseréje. Az adathalmaz struktúrájának tervezése és implementációja, a végső adat ellenőrzése. Az adathalmaz vizsgálata, statisztikai adatok mérése.
    - András Kicsi, László Vidács, and Tibor Gyimothy. Testroutes: A manually curated method level dataset for test-to-code traceability. In Proceedings of the 17th International Conference on Mining Software Repositories, MSR 2020, pages 593–597. IEEE, IEEE, jun 2020.
7. Névkonvenciók elemzésének tervezése, különböző névkonvenciós módszerek kidolgozása, kinyerésük implementációja. Névkonvenciós eredmények vizsgálata, értelmezése a kód áttekintésével, következtetések levonása. Kombinált megoldások

eredményeinek vizsgálata és magyarázata. A TestRoutes adathalmazon történő kiértékelés.

- András Kicsi, Viktor Csuvik, and László Vidács. Large Scale Evaluation of NLP-based Test-to-Code Traceability Approaches. IEEE Access, 2021.

A **harmadik tézisponthoz** (*Machine Understanding of Radiologic Reports*) tartozó eredmények:

1. Az annotációs szisztéma alapjainak kidolgozása (testrészek, elváltozások, tulajdonságok, ezek konvenciói), az annotációs szisztéma megállapításának és finomításának összegzése és véglegesítése. Kommunikáció az annotátorokkal, rendszeres felügyelet és segítség az annotáció kérdéses eseteiben. Annotáció előkészítése, annotációs útmutató készítése. A tanítás során felhasznált adatok tisztítása, összegzése, egyeztetése. Eredmények értelmezése, magyarázata, kérdéses esetek javítása a tanítóadatokban.
    - András Kicsi, Péter Pusztai, Klaudia Szabó Ledenyi, Endre Szabó, Gábor Berend, Veronika Vincze, and László Vidács. Információkinyerés magyar nyelvű gerinc MR leletekből. In XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019), page 177–186, Szeged, 2019. (Magyar nyelven)
    - András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, Péter Németh, and László Vidács. Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 15–28, Szeged, 2020. (Magyar nyelven)
    - András Kicsi, Klaudia Szabó Ledenyi, Péter Németh, Péter Pusztai, László Vidács, and Tibor Gyimóthy. Elírások automatikus detektálása és javítása radiológiai leletek szövegében. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 191–204, Szeged, 2020. (Magyar nyelven)
    - András Kicsi, Péter Pusztai, Endre Szabó, and László Vidács. Szaknyelvi annotációk javításának statisztikai alapú támogatása. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 115–128, Szeged, 2020. (Magyar nyelven)
    - András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, and László Vidács. Automatic classification and entity relation detection in hungarian spinal MRI reports. In 3rd ICSE Workshop on Software Engineering for Healthcare, 2021.
2. A kapcsolatok szabály-alapú felismerésének alapjainak megtervezése, az kezdetleges ontológián alapuló azonosítás alapjainak megtervezése.
    - András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, Péter Németh, and László Vidács. Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 15–28, Szeged, 2020. (Magyar nyelven)
    - András Kicsi, Klaudia Szabó Ledenyi, Péter Németh, Péter Pusztai, László Vidács, and Tibor Gyimóthy. Elírások automatikus detektálása és javítása radiológiai leletek szövegében. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 191–204, Szeged, 2020. (Magyar nyelven)
3. Tagmondat-alapú tagadáskezelés megtervezése.

- András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, Péter Németh, and László Vidács. Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 15–28, Szeged, 2020. (Magyar nyelven)
- András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, and László Vidács. Automatic classification and entity relation detection in hungarian spinal MRI reports. In 3rd ICSE Workshop on Software Engineering for Healthcare, 2021.

4. Keresőfelület megtervezése, ehhez adatbázis tervezése.
    - András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, Péter Németh, and László Vidács. Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 15–28, Szeged, 2020. (Magyar nyelven)
5. Helyesírásjavítási kísérletek koordinációja, ebben kézi annotátor instrukciója és felügyelete, kérdéses esetekben segítség nyújtása. Helyes testrész-szótár elkészítése a leletekből válogatott szavak alapján. Eredmények kézi ellenőrzése, egyeztetése az annotátorral. Prototípus-felület működésének tervezése.
    - András Kicsi, Klaudia Szabó Ledenyi, Péter Németh, Péter Pusztai, László Vidács, and Tibor Gyimóthy. Elírások automatikus detektálása és javítása radiológiai leletek szövegében. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 191–204, Szeged, 2020. (Magyar nyelven)
6. Brat-ban történő statisztika-megjelenítés koncepciója. A statisztikai adatokat kommentként beszúró és egy gyors keresést segítő szkript implementációja az annotáció javításhoz. Annotáció javítás leíró dokumentuma, kapcsolattartás az annotátorral, és segítség nyújtása az annotáció során.
    - András Kicsi, Péter Pusztai, Endre Szabó, and László Vidács. Szaknyelvi annotációk javításának statisztikai alapú támogatása. In XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), page 115–128, Szeged, 2020. (Magyar nyelven)
7. Módszer kidolgozása leletekből megismert információ automatikus vizualizációjára.
    - András Kicsi, Klaudia Szabó Ledenyi, Péter Pusztai, and László Vidács. Automatic classification and entity relation detection in hungarian spinal MRI reports. In 3rd ICSE Workshop on Software Engineering for Healthcare, 2021.

Ezek az eredmények **Kicsi András** PhD disszertációján kívül más tudományos fokozat megszerzésére nem használhatók fel.

Szeged, 2022.08.31

Kicsi András

jelölt

Dr. Vidács László

témavezető

Az Informatika Doktori Iskola vezetője kijelenti, hogy jelen nyilatkozatot minden társszerzőhöz eljuttatta, és azzal szemben egyetlen társszerző sem emelt kifogást.

Szeged, 2022. 09. 12.

SZEGEDI TUDOMÁNYEGYETEM
INFORMATIKA DOKTORI ISKOLA
SZEGED

Dr. Jelasity Márk

Informatikai Doktori Iskola vezetője