Stefanie Brüninghaus and Kevin D. Ashley (2005) Reasoning with Textual Cases. In: Munoz-Aliva, H. and Ricci, F. (Editors). Case-Based Reasoning Research and Development: Proceedings of the 4th. International Conference on Case-Based Reasoning (ICCBR-05). Chicago, IL. August 2005. Springer Verlag: Heidelberg, Germany. Lecture Notes in Artificial Intelligence LNAI 3620. Pages 137-151.

Reasoning with Textual Cases

Stefanie Brüninghaus and Kevin D. Ashley

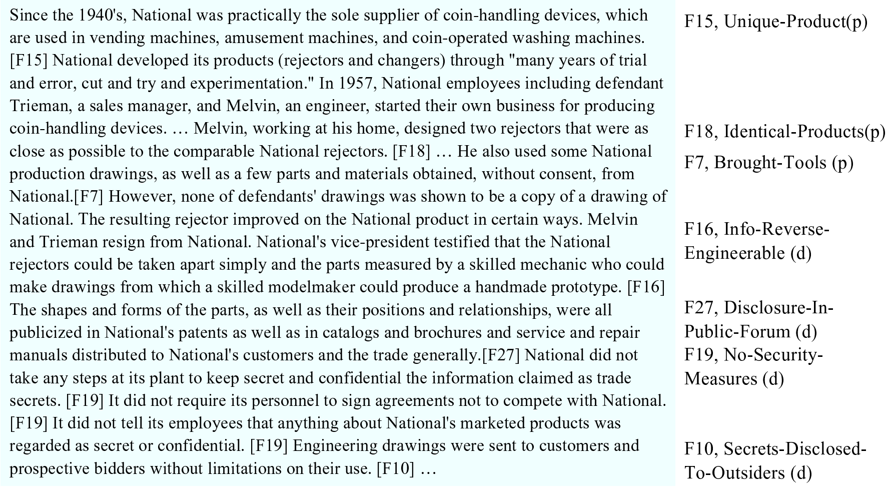
Learning Research and Development Center,
Intelligent Systems Program, and School of Law
University of Pittsburgh
3939 O'Hara Street; Pittsburgh, PA 15260 (USA)

Abstract. This paper presents methods that support automatically finding abstract indexing concepts in textual cases and demonstrates how these cases can be used in an interpretive CBR system to carry out case-based argumentation and prediction from text cases. We implemented and evaluated these methods in SMILE+IBP, which predicts the outcome of legal cases given a textual summary. Our approach uses classification-based methods for assigning indices. In our experiments, we compare different methods for representing text cases, and also consider multiple learning algorithms. The evaluation shows that a text representation that combines some background knowledge and NLP combined with a nearest neighbor algorithm leads to the best performance for our TCBR task.

1 Introduction

The goal of researchers investigating Textual CBR (TCBR) has been to enable traditional CBR systems to deal directly and intelligently with cases described as text. So far, the work has focused on retrieving cases to help a human solve textually-described problems, or on assigning indices to or highlighting passages in textual cases humans will use in solving problems. An essential difference between TCBR and Information Retrieval (IR) has been that IR methods tend not to take into account much semantic information or background domain knowledge about problem-solving. By contrast, TCBR methods attempt to leverage domain knowledge. Their indexing and retrieval mechanisms apply domain-specific, problem-solving knowledge, as well as more general knowledge, to process texts to help readers solve specific problems (Lenz 1999, p. 298; Burke 1998, pp. 13-14). IR researchers tend to dismiss such domain-specific techniques as *ad hoc*; textual CBR systems "eschew flexibility and generality for precision and utility for a given group of users." (Burke 1998).

In the relatively brief history of TCBR research, it is a human reasoning agent who solves the textually described problem using the cases returned by the program. However, in the work reported here, an automated reasoning agent solves the problems inputted as texts. Specifically, we describe a program called SMILE+IBP that uses CBR to predict the outcomes of legal disputes inputted directly as text and to explain those predictions. Fig. 1 shows an example of an annotated case text, the squib summarizing the facts of National Rejectors v. Trieman, 409 S.W.2d 1 (Mo.1966), and Fig. 2 shows SMILE+IBP's output for this case. The inputs to its predictive component, the Issue-Based Prediction program IBP, are representations of the problem facts in terms



Since the 1940's, National was practically the sole supplier of coin-handling devices, which are used in vending machines, amusement machines, and coin-operated washing machines. [F15] National developed its products (rejectors and changers) through "many years of trial and error, cut and try and experimentation." In 1957, National employees including defendant Trieman, a sales manager, and Melvin, an engineer, started their own business for producing coin-handling devices. … Melvin, working at his home, designed two rejectors that were as close as possible to the comparable National rejectors. [F18] … He also used some National production drawings, as well as a few parts and materials obtained, without consent, from National.[F7] However, none of defendants' drawings was shown to be a copy of a drawing of National. The resulting rejector improved on the National product in certain ways. Melvin and Trieman resign from National. National's vice-president testified that the National rejectors could be taken apart simply and the parts measured by a skilled mechanic who could make drawings from which a skilled modelmaker could produce a handmade prototype. [F16] The shapes and forms of the parts, as well as their positions and relationships, were all publicized in National's patents as well as in catalogs and brochures and service and repair manuals distributed to National's customers and the trade generally.[F27] National did not take any steps at its plant to keep secret and confidential the information claimed as trade secrets. [F19] It did not require its personnel to sign agreements not to compete with National. [F19] It did not tell its employees that anything about National's marketed products was regarded as secret or confidential. [F19] Engineering drawings were sent to customers and prospective bidders without limitations on their use. [F10] …

F15, Unique-Product(p)

F18, Identical-Products(p)

F7, Brought-Tools (p)

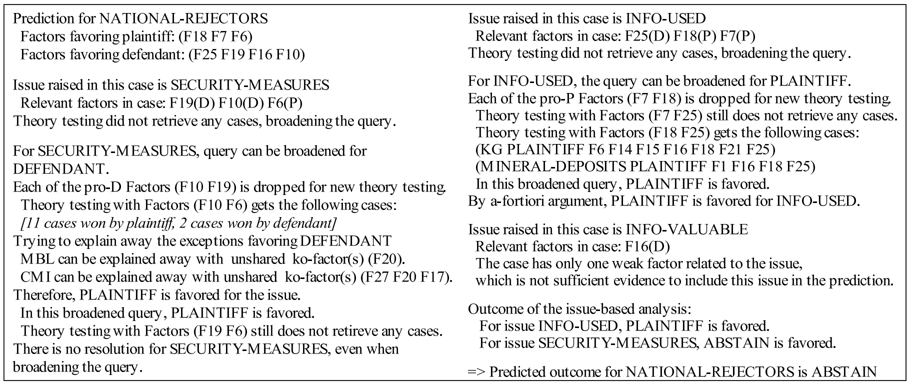
F16, Info-Reverse-Engineerable (d)

F27, Disclosure-In-Public-Forum (d)

F19, No-Security-Measures (d)

F10, Secrets-Disclosed-To-Outsiders (d)

Fig. 1. Summary of the *National Rejectors* case, annotated with applicable Factors



```
Prediction for NATIONAL-REJECTORS
  Factors favoring plaintiff: (F18 F7 F6)
  Factors favoring defendant: (F25 F19 F16 F10)

Issue raised in this case is SECURITY-MEASURES
  Relevant factors in case: F19(D) F10(D) F6(P)
Theory testing did not retrieve any cases, broadening the query.

For SECURITY-MEASURES, query can be broadened for
DEFENDANT.
Each of the pro-D Factors (F10 F19) is dropped for new theory testing.
  Theory testing with Factors (F10 F6) gets the following cases:
  [11 cases won by plaintiff, 2 cases won by defendant]
Trying to explain away the exceptions favoring DEFENDANT
  MBL can be explained away with  unshared  ko-factor(s) (F20).
  CMI can be explained away with unshared  ko-factor(s) (F27 F20 F17).
Therefore, PLAINTIFF is favored for the issue.
  In this broadened query, PLAINTIFF is favored.
  Theory testing with Factors (F19 F6) still does not retireve any cases.
There is no resolution for SECURITY-MEASURES, even when
broadening the query.

Issue raised in this case is INFO-USED
  Relevant factors in case: F25(D) F18(P) F7(P)
Theory testing did not retrieve any cases, broadening the query.

For INFO-USED, the query can be broadened for PLAINTIFF.
Each of the pro-P Factors (F7 F18) is dropped for new theory testing.
  Theory testing with Factors (F7 F25) still does not retrieve any cases.
  Theory testing with Factors (F18 F25) gets the following cases:
  (KG PLAINTIFF F6 F14 F15 F16 F18 F21 F25)
  (MINERAL-DEPOSITS PLAINTIFF F1 F16 F18 F25)
  In this broadened query, PLAINTIFF is favored.
By a-fortiori argument, PLAINTIFF is favored for INFO-USED.

Issue raised in this case is INFO-VALUABLE
  Relevant factors in case: F16(D)
  The case has only one weak factor related to the issue,
  which is not sufficient evidence to include this issue in the prediction.

Outcome of the issue-based analysis:
   For issue INFO-USED, PLAINTIFF is favored.
   For issue SECURITY-MEASURES, ABSTAIN is favored.

=> Predicted outcome for NATIONAL-REJECTORS is ABSTAIN
```

Fig. 2. Case-based analysis of *National Rejectors* text by SMILE+IBP

of abstract features, called Factors. These are prototypical fact patterns that tend to favor plaintiff's (p) or defendant's (d) position (Aleven 2003; Ashley 1990). The classification component, SMILE (SMart Index LEarner) assigns these features automatically to the textual description of the problem's facts using classifiers learned from a database of marked-up case texts. This integration of IBP and SMILE, as shown in Fig.3, allows us to assess the quality of SMILE's index assignments and particularly to test two hypotheses about the best way to represent case texts for learning classifiers. The text representation techniques are alternative means for capturing the kind of domain-specific, problem-solving knowledge and more general knowledge that enable a traditional CBR system to process case texts. In this way, we use enhanced text representations and machine learning to make TCBR techniques more general and automatic while preserving their focus on domain-specific problem-solving.

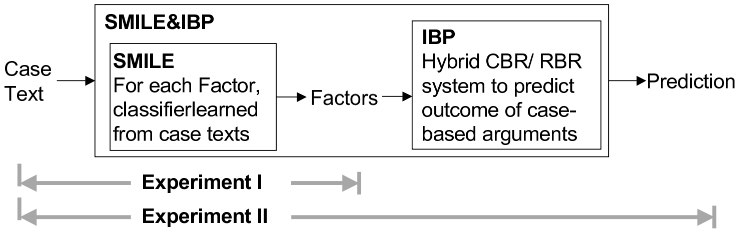

SMILE&IBP
Case Text
SMILE
For each Factor, classifierlearned from case texts
Factors
IBP
Hybrid CBR/ RBR system to predict outcome of case-based arguments
Prediction
Experiment I
Experiment II

Fig. 3. Setup of the SMILE+IBP system, and outline of the experiments

2 Text Representation for TCBR

The most widely used text representation in TCBR has been a bag-of-words, in which the text is tokenized into single words, thereby doing away with word order. One of the first projects aimed at indexing textual cases, SPIRE (Daniels & Rissland 1997), used a small collection of excerpts related to its indexing concepts to locate the most promising text passages in a new unseen text. SPIRE relied on the passage retrieval module of an IR system, which represents texts as a bag-of-words, to find those sections in a new case that are most similar to its sample excerpts. The experiments compared different weighting schemes and variations of a bag-of-words representation. Other TCBR projects focused on the retrieval of text cases, rather than assigning indices. (Burke *et al.* 1997) and (Lenz 1999) showed that adding semantic information from WordNet can lead to better performance in retrieval-oriented TCBR systems. Recent work in TCBR has considered other, more advanced representations. (Cunningham *et al.* 2004) present a promising approach, which maintains some syntactic information by translating text into a network structure. An evaluation in the legal domain remains somewhat inconclusive, and further experiments will be necessary to show whether this intuitively appealing approach will lead to better performance. SCALIR was developed before the term TCBR was introduced. It also relied on representing legal cases texts in a network structure, with favorable results (Rose 1994). A promising and highly ambitious approach, using natural language processing (NLP) to derive a deep, logical representation, has been proposed for the FACIT project (Gupta & Aha 2004).

As this overview suggests, representation remains a central issue for TCBR; many researchers are exploring better representations for text cases. Our research carries this a step further, in that we incorporate shallow NLP, and in that our CBR application actually reasons with the automatically indexed cases.

Our approach to representing text cases was motivated by three observations and intuitions we gained from indexing text cases manually. First, our collection of legal cases comprises cases from many different jurisdictions and procedural settings, covering a period of about 50 years. This variety is reflected in the texts. The authors follow different stylistic conventions and often use a different vocabulary, as well. For instance, some judges, especially in older cases, tend to use “covenant,” whereas others prefer the terms “contract” or “agreement.” Adding some form of semantic knowledge to a lexicon may help an indexing program to find commonalities between examples that use a different vocabulary. In past experiments (Brüninghaus & Ashley 1999), we found that adding a thesaurus can lead to performance improvements.

Second, the names of the parties involved in a lawsuit are of little use for indexing, especially in long and complex cases. Keeping track of different names can be hard for humans, and is beyond today's computer systems. Instead, replacing names by their roles in the case makes cases more readable and enables a learning algorithm to better generalize from cases. Moreover, the same name can occur in different cases, sometimes even in different roles. Our collection has two cases involving IBM as plaintiff. Any inferences based on the name IBM would be erroneous, however, because the cases are completely unrelated and involve different scenarios. We hypothesize that replacing the names of parties and products by their roles in the case will lead to better indexing because it allows learning algorithms to generalize more accurately from examples.

Third, word order and other syntactic features are crucial for assigning some more complex indexing concepts. Consider "Plaintiff sent a letter to defendant," which often is related to Factor F1, Disclosure-In-Negotiations (d). A sentence with almost the same words, "A letter was sent to plaintiff by defendant," is not related to F1. In order to distinguish these two instances, at least some level of linguistic analysis and representation of syntactic features are required, like passive voice and the relations between constituents.

More formally, these intuitions inspired the following research hypotheses:

Hypothesis I Abstracting from names and individual entities in a case text to their roles in the case allows a learning algorithm to better generalize from training examples.

Hypothesis II Using some linguistic analysis to capture (1) patterns of actions and (2) negation preserves crucial information from the text and thereby leads to better classification.

In order to test our intuitions empirically, we implemented two representations that correspond to adding the knowledge as per the above hypotheses, as well as a baseline.

Bag-of-words/BOW Our baseline representation, against which all measures will be compared, is the basic bag-of-words. The text is tokenized into single words, whose relative position to each other is lost in the process. We do not eliminate stopwords, and we do not remove suffixes with a stemmer. For instance, consider this sentence from the *ICM* case, which is evidence for Factor F7, Brought-Tools (p): "Newlin copied some files from ICM and brought them with him to DTI." In BOW, it would be represented as AND BROUGHT COPIED DTI FILE FROM HIM ICM NEWLIN SOME THEM TO WITH.

Roles-Replaced/RR In this representation, names and references to individual entities in the text are replaced by their roles in the lawsuit. The sentence above would become "Defendant copied some information from plaintiff and brought them with him to defendant." Then, this example is tokenized as a bag-of-words, AND BROUGHT COPIED DEFENDANT HIM INFORMATION PLAINTIFF SOME THEM TO WITH. While this representation is still limited to the degree that it is a bag-of-words, it contains more relevant information about the case facts than the previous representation as BOW.

For our evaluation, we assumed there is a program that can automatically replace names by roles with high accuracy, as suggested in (Brüninghaus & Ashley 2001). In the experiments, we relied on texts where this substitution had been carried out manually by experts. With error-free replacements, all observed differences in performance, or their absence, could be attributed to the representation, rather than to an implementation for role replacements.

Propositional-Patterns/ProP Propositional patterns are intended to capture more of the meaning contained in a sentence and thereby overcome some of the problems of a bag-of-words (Brüninghaus & Ashley 2001). They are powerful features and combine two terms that are in a specified syntactic relation with each other. ProPs differ from bigrams, pairs of adjacent words sometimes used as features in IR, in that we use syntax, and not adjacency, as a criterion. ProPs were inspired by the automatically generated caseframes in the AutoSlog-TS system (Riloff 2003).For SMILE, we are using Ellen Riloff's AutoSlog tools, which include Sundance, a robust partial parser that can be easily configured and adapted for new domains.

Roughly speaking, ProPs combine the headword of the trigger, the most relevant word of the "if" part, and headword of the filler, the most relevant word of the "then" part, of the extraction rules in an IE system. In addition to the syntactic knowledge, ProPs also capture some of the semantic knowledge from Sundance's lexicon, similar to the integration of a thesaurus presented in (Brüninghaus & Ashley 1999), by adding a new ProP for each synonym of the constituent words. Thus, for the sentence from *ICM*, one would get (DEFENDANT COPY) (PERSON COPY) (COPY INFORMATION) (COPY_FROM PERSON) (COPY_FROM PLAINTIFF) (DEFENDANT BRING) (PERSON BRING) (BRING THEM) (BRING_TO DEFENDANT) (BRING_TO PERSON) (BRING_WITH HIM). While this representation is still fairly simple, it is much more likely to allow the inference that Factor F7 applies than the RR representation. In our experiments, this sentence was correctly classified as F7 by RR and ProP, but not by BOW.

Generating the RR representation from the original text corresponds to Hypothesis I, replacing names by roles. We therefore expect that, according to Hypothesis I, the results with cases represented as RR in our experiments will be better than with BOW, or RR $>$ BOW. Deriving ProPs from text where the names are replaced by roles corresponds to Hypothesis II. Consequently, if Hypothesis II applies, ProP $>$ RR. Since ProPs are derived from text in which names are replaced by roles, we also expect that the Hypotheses are transitive and that ProP $>$ BOW.

3 Integration of Indexing and Reasoning in SMILE+IBP

We tested these hypotheses in the context of our SMILE (Brüninghaus & Ashley 2001) and IBP (Brüninghaus & Ashley 2003) programs.

3.1 Classification-Based Indexing in SMILE

For the task of assigning indices, our case base with manually indexed cases together with the textual representation of these cases can be viewed as a set of examples for "how indexing should be done." Following this characterization of the problem, we take a classification-based approach to indexing in SMILE, treating our existing case base as training set, and the Factors as target concepts.

As a machine learning (ML) approach, SMILE has two phases, classification and training; see Fig. 4. In the classification phase, SMILE works in a very modular way. It has 26 separate classifiers, one for each Factor F1 to F27 (for historical reasons, there is no F9).[1] Unlike many other text learning approaches, we treat the text cases as a set of example sentences, rather than one example document. Evidence for Factors is usually found in sentences, as illustrated in the *National Rejectors* squib in Fig. 1. SMILE first splits a new case, which does not have any mark-ups, into a set of sentences and represents them as BOW, RR or ProP. These sentences are then given as input to each of the classifiers. SMILE assigns a Factor if at least one sentence from the case text is labeled as a positive instance. The applicable Factors from all classifiers are then collected for SMILE's output.

In the training phase, the training set consists of the squibs from our collection, marked up with the applicable Factors similar to the *National Rejectors* squib in Fig. 1. As noted, SMILE learns separate classifiers for each Factor. It takes the cases where the Factor applies, and collects the sentences marked up with the Factor as positive training examples. All other sentences, those not marked up from a case where the Factor applies as well as the sentences from the cases without the Factor, are collected as negative training examples. The training examples are represented as BOW, RR, or ProP, and given as inputs to the learning algorithm. The learned classifiers are used as illustrated in Fig. 4.

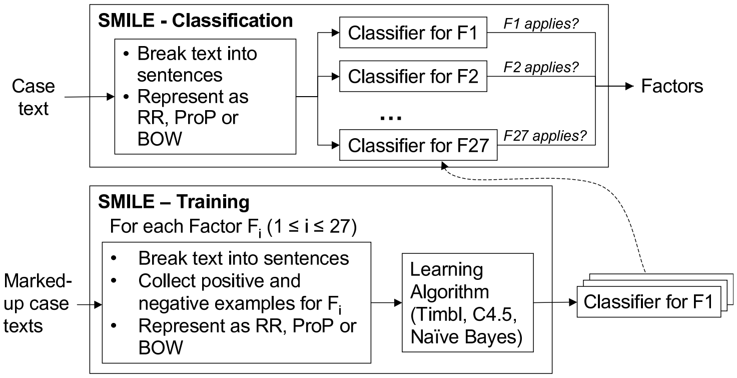

SMILE - Classification
Case text
• Break text into sentences
• Represent as RR, ProP or BOW
Classifier for F1
Classifier for F2
...
Classifier for F27
F1 applies?
F2 applies?
F27 applies?
Factors
SMILE – Training
For each Factor F_i ($1 \leq i \leq 27$)
Marked-up case texts
• Break text into sentences
• Collect positive and negative examples for F_i
• Represent as RR, ProP or BOW
Learning Algorithm (Timbl, C4.5, Naïve Bayes)
Classifier for F1

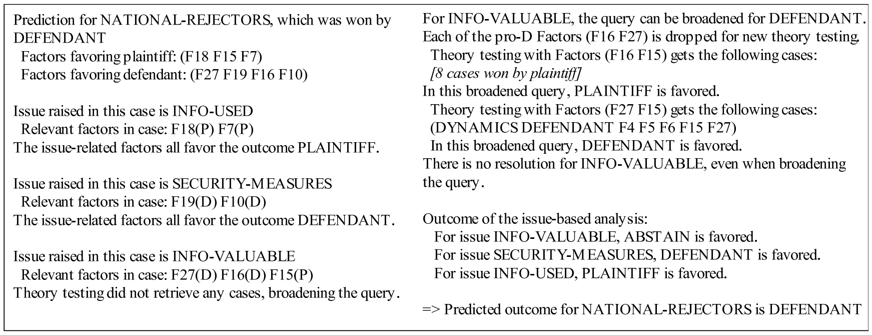
Fig. 4. Architecture of the SMILE system

3.2 Issue-Based Prediction in IBP

IBP is hybrid case-based/rule-based algorithm that predicts the outcome of legal cases. Due to space limitations, we can only give a brief description of IBP, it is discussed in detail in (Brüninghaus & Ashley 2003) and (Ashley & Brüninghaus 2003). IBP combines a weakly-predictive domain model, which was derived from authoritative legal

[1] Strictly speaking, although NN implements classification based on past cases, it does not explicitly learn a classifier in the sense that the other ML approaches do. For this paper, we ignore that difference, and treat NN like other ML algorithms.

sources, and a CBR module. The domain model captures the general structure of the domain, the issues and their relations. It relates Factors to issues, but does not include rules to resolve conflicting Factors. For instance, the domain model captures that plaintiff has to show that the information was a trade secret in order to win a claim for trade secret misappropriation. This requires that the information was valuable and that measures were taken to keep it a secret. Based on Hypo, IBP's CBR module supports a form of scientific hypothesis testing to resolve conflicting evidence related to an issue.



```
Prediction for NATIONAL-REJECTORS, which was won by
DEFENDANT
  Factors favoring plaintiff: (F18 F15 F7)
  Factors favoring defendant: (F27 F19 F16 F10)

Issue raised in this case is INFO-USED
  Relevant factors in case: F18(P) F7(P)
The issue-related factors all favor the outcome PLAINTIFF.

Issue raised in this case is SECURITY-MEASURES
  Relevant factors in case: F19(D) F10(D)
The issue-related factors all favor the outcome DEFENDANT.

Issue raised in this case is INFO-VALUABLE
  Relevant factors in case: F27(D) F16(D) F15(P)
Theory testing did not retrieve any cases, broadening the query.
For INFO-VALUABLE, the query can be broadened for DEFENDANT.
Each of the pro-D Factors (F16 F27) is dropped for new theory testing.
  Theory testing with Factors (F16 F15) gets the following cases:
  [8 cases won by plaintiff]
In this broadened query, PLAINTIFF is favored.
  Theory testing with Factors (F27 F15) gets the following cases:
  (DYNAMICS DEFENDANT F4 F5 F6 F15 F27)
  In this broadened query, DEFENDANT is favored.
There is no resolution for INFO-VALUABLE, even when broadening
the query.

Outcome of the issue-based analysis:
   For issue INFO-VALUABLE, ABSTAIN is favored.
   For issue SECURITY-MEASURES, DEFENDANT is favored.
   For issue INFO-USED, PLAINTIFF is favored.

=> Predicted outcome for NATIONAL-REJECTORS is DEFENDANT
```

Fig. 5. IBP's analysis of *National Rejectors*, Factors manually assigned by an expert

Fig. 5 shows IBP's analysis, given an expert's manual interpretation of the *National Rejectors* case. Factors F27, Disclosure-In-Public-Forum (d), F16, Info-Reverse-Engineerable (d) and F15, Unique-Product (p) are related to the issue whether the information is valuable. IBP can conclude that the issue is raised, but needs to rely on its CBR module to find which side is favored on the issue. If the CBR module fails to resolve conflicting evidence, as in *National Rejectors*, IBP abstains on the issue.

We evaluated IBP and compared its predictions to several other ML and CBR-based methods. We found that IBP's predictions are most accurate, with a significant margin (Brüninghaus & Ashley 2003).

3.3 Combination of Indexing and Prediction

Our approaches to indexing in SMILE and prediction in IBP can be combined into SMILE+IBP by using the Factors assigned by SMILE as input to IBP as illustrated in Fig. 3. Thereby, we are in a position to generate a case-based analysis and prediction for cases from case texts, without manual intervention beyond converting and copying files. By combining SMILE and IBP, we have created a TCBR system that can carry out real reasoning beyond just retrieval.

4 Evaluation

Using SMILE and SMILE+IBP, we ran a set of experiments to test the above hypotheses to find out what makes a good text representation. We tried different representations for the cases to be indexed, and measured performance for Factor assignments as well as prediction using the automatically assigned Factors.

4.1 Experimental Design

In these experiments, we used 146 cases from the CATO case database (Aleven 2003) in two forms, represented as a set of applicable Factors and the squibs. The cases represented as Factors were used as the case base for IBP. The squibs summarize the courts' written opinions and help students infer which Factors apply in a case. The full-text opinions tend to be fairly long; *National Rejectors'* is 48 pages. In writing the squibs, the authors were encouraged to copy-and-paste from the opinions' descriptions of case facts. Only a relatively small part of the text of the squibs was written from scratch. The squibs were manually marked up for inclusion in SMILE.

While the manual mark-up for SMILE usually corresponds to CATO's list of Factors, there are some differences. For SMILE, we require that the evidence for a Factor is explicit in the text, and that no indirect inferences are needed even if they are based on common-sense interpretations of the text. As a result, some examples of the harder-to-find Factors were not included in the mark-up, especially Factors F3, Employee-Sole-Developer (d) and F5, Agreement-Not-Specific (d). We also decided not to follow CATO's conventions for Factors F6, Security-Measures (p) and F19, No-Security-Measures (d). In CATO's cases, Factor F6 is assigned whenever there are any security measures. F19 will not apply, even when there is overwhelming evidence that the plaintiff neglected security measures as long as it took some measures. For SMILE, however, F19 is marked up whenever the court explicitly focuses on instances where the plaintiff neglected to take certain security measures, even if plaintiff took some other measures.

The experiments were conducted as a leave-one-out cross-validation over all cases in the collection. For instance, when the *National Rejectors* case was the test example, its squib was included neither in the training set for SMILE, nor in IBP's database for testing predictions; in this run, *National Rejectors* played no role in training SMILE's classifiers or in IBP's predictions.

While the focus of our work is primarily on finding the best text representation, we included three learning algorithms with very different characteristics and learning biases. We considered multiple algorithms because it was not clear *a priori* how suitable these algorithms would be for our task. These algorithms are commonly used in text classification experiments and include Nearest Neighbor (NN), Decision Trees, and Naive Bayes. We selected respectively: Timbl (Daelemans *et al.* 2004), C4.5 (Quinlan 2004) and Rainbow (McCallum 2004). All are suitable for learning from text, freely available implementations from reliable sources. We used default parameters for Timbl, which in particular means k = 1 (i.e., 1-NN). We explored other parameter settings, but found that 1-NN was preferable. In C4.5, we set pruning to 100% confidence level.

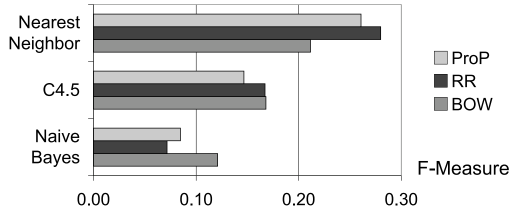
Our experiments were run for three algorithms, three representations, 146 iterations of cross-validation, and 26 Factors, for an overall of about 35,000 experiment runs. The data included about 2,000 example sentences, with about 2,000 features for each representation. The experiments for this in-depth evaluation on rather complex data ran around the clock for several weeks. After our experiments, the most suitable representation and algorithm can be identified to learn one classifier for each Factor, which will be more efficient by three orders of magnitude.

In analyzing the results, we applied statistical tests to find whether the observed differences are statistically significant, or merely caused by random effects. Because our experiments were run as cross-validation, the commonly used T-test may not lead to reliable results (Dietterich 1996; Salzberg 1997). Based on the recommendations in (Dietterich 1996), we used Wilcoxon's Signed-Rank test, a so-called non-parametric test. A common pitfall in comparing multiple algorithms (or in our case, representations) is the repeated, pairwise comparison of results. However, this requires that a significant difference among all alternatives is shown first. We used Friedman's test for this purpose. Following convention, we say that results with $p < 0.05$ are statistically significant (Cohen 1995).

4.2 Experiment I

In our first set of experiments, we compared the effect of representation and learning algorithm on Factor assignment. We kept everything fixed; the only change was the combination of learning algorithm and representation. As a result, all observed differences can be attributed to these conditions.

We followed the evaluation commonly carried out for text classification. As illustrated in Fig. 3, the input in Experiment I is a raw case text, without any annotations, the output a set of Factors. Performance for each Factor was measured in terms of the F-measure, which is defined as the harmonic mean of precision and recall (Cohen 1995) as follows: $F = \frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$.



F-Measure	ProP	RR	BOW
Nearest Neighbor	0.261	0.280	0.211
C4.5	0. 147	0.167	0. 168
Naïve Bayes	0.085	0.072	0.121

Fig. 6. Average F-measure for Experiment I

The averaged results over all Factors in Fig. 6 show two major results: Timbl is the best learning algorithm for the task, and RR and ProP outperform BOW.

First, let us consider the differences between the representations, focusing on the best algorithm in the experiments, the Timbl implementation of NN. Friedman's test indicates that the differences among all nine variants are statistically significant. For the results with Timbl, in the top of both the chart and the table in Fig. 6, Wilcoxon's Ranked-Sign test shows that scores for ProP and RR are significantly higher that BOW. These results provide evidence for Hypothesis I. RR has a higher score than ProP, which is not consistent with Hypothesis II; however, the difference is not statistically significant. The second experiment discussed in Section 4.3 provides additional evidence concerning Hypothesis II.

One reason that Hypothesis II could not be confirmed here is that NLP remains one of the main problems with generating ProPs. Even though Sundance is a very robust state-of-the-art parser, its outputs are not always accurate enough for high-performance indexing from complex texts. In *National Rejectors*, ProP does not find Factor F27, Disclosure-in-Public-Forum (d). The relevant sentence is "The shapes and forms of the parts, as well as their positions and relationships, were all publicized in plaintiff's patents as well as in catalogs and brochures and service and repair manuals distributed to plaintiff's customers and the trade generally." This sentence has several constructs that are notoriously difficult to parse. In the first clause, Sundance gets confused by the verb phrase "were all publicized," and parses it as an active verb construction, with "all" as the subject. As a consequence, the ProPs generated for the sentence are hardly of any use for assigning indices, and Factor F27 is missed. In order to show that the problem is not a general limitation of ProPs, but rather caused by language too complex for the parser, we modified the grammar to "The shapes and forms of the parts and their positions and relationships were publicized in plaintiff's patents, catalogs and brochures and manuals, which were distributed to plaintiff's customers and the general trade." When we manually added the ProPs for this sentence's parse, it was correctly classified as an instance of F27. The sentence retrieved as most similar comes from the *Dynamics* case: "The first two of these features were publicized in a conference paper and an advertizing brochure." This example shows how minor grammatical adjustments can lead to correct Factor assignments. It indicates that our experiments are a lower bound on the performance of ProPs, which most likely will increase with more accurate parsing.

When we focus on the relative strengths of ProP and RR, we find that RR tends to have an advantage for Factors that favor plaintiff, while ProP tends to have an advantage for Factors that favor defendant. It appears that a number of pro-plaintiff Factors capture situations or features of the product, like Factor F15, Unique-Product (p). Such Factors can be represented fairly well through single words, and thus RR often suffices. Several of the pro-defendant Factors, on the other hand, describe defendant's actions, like F27. This requires more information about "who did what" and lends itself to representation with ProPs. Experiment II will further investigate how these relative strengths of ProP and RR have an impact on reasoning with the cases.

Second, with regard to the best learning algorithm, we found that NN outperforms C4.5 and Naive Bayes. The latter is remarkable because Naive Bayes is often hard to beat for text classification tasks. In our experiments, Naive Bayes had fairly good scores for only one Factor, F6, Security-Measures(p), which has 198 sentences marked-up. For most other Factors, it failed to find any instances.

In our collection, it appears that the conditions are not favorable for Naive Bayes. The distributions are extremely skewed. We have around 2,000 example sentences, yet, for some Factors, fewer than ten sentences are marked up. Thus, the prior probability for a Factor is low. In addition, the relevant vocabulary is large, around 2,000 features for each representation. With such sparse data, there may not be sufficiently many examples to derive reliable probability estimates.

On the other hand, NN does not rely on the prior class probabilities and has an advantage, especially for Factors with very few positive instances. Our experience confirms the reasoning of (Cardie & Howe 1997), who had good results with a NN approach for an information extraction task. They chose NN for an application where the distributions are highly skewed and where the goal is to find the minority class.

Similarly, the experiments show that C4.5 is not ideal for the task. The inductive bias of C4.5 is to learn trees with short branches; it does best when it can find relatively few highly predictive words. However, C4.5 is less suited for more complex concepts, or Factors, where multiple features may have to be weighed in a context-sensitive manner. Another problem is the small number of positive instances for many of the Factors. For instance, apart from the *National Rejectors* case, only *Dynamics* uses "was publicized" in relation to F27. Thus, it would not be possible for C4.5 to correctly find F27 in *National Rejectors*. In general, C4.5 has no way to generalize from singletons. A NN approach, on the other hand, is more suitable for such concepts. Moreover, C4.5 appears to provide evidence against Hypothesis II (see the middle column and line in Fig. 6); the results for ProP are much lower than both BOW and RR. However, this observation is related to the fact that C4.5 is not appropriate for singleton examples. ProPs are a more powerful, but also more specific representation than BOW and RR, which alleviates the problem of rare examples.

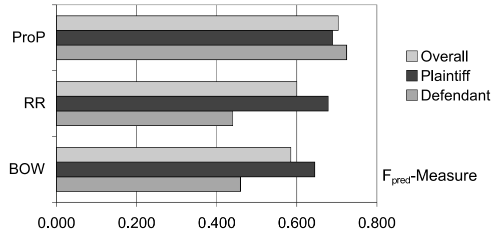
4.3 Experiment II

While Experiment I gives some important insights into how different representations have an impact assigning individual Factors, it only considers Factors in isolation. It does not capture the overall gestalt of a case, the interactions among Factors, or how some Factors are more relevant for our CBR task than others. For instance, not assigning Factor F27 to *National Rejectors*, as discussed above, is a critical error that can have a strong impact on prediction. In Experiment II, we therefore push a step beyond the sort of evaluation commonly carried out for text classification experiments by including prediction as an indicator for how well the assigned Factors capture the contents of a case. As before, the inputs in Experiment II are the squibs, but the outputs are the predictions of the cases' outcomes, as illustrated in Fig. 3. As in Experiment I, we kept everything else constant; the only difference was the representation of the text cases. Thus, all observed differences in performance can be attributed to the representation.

The experiments were scored by comparing IBP's predictions to the cases' real outcomes. We recorded accuracy over the cases where IBP made a prediction, as well as coverage, which is the percentage of cases where IBP made a prediction. Then, we combined these by adapting the F-measure for predictions: $\mathrm{F}_{\mathrm{pred}} = \frac{2*\mathrm{accuracy}*\mathrm{coverage}}{\mathrm{accuracy}+\mathrm{coverage}}$.

ProP has the best performance, with an $\mathrm{F}_{\mathrm{pred}}$-measure of 0.703, followed by RR, with 0.6 and BOW with 0.585; see "Overall" in Fig. 7. The difference among the representations is statistically significant, using Friedman's test. The difference between ProP and RR, as well as between ProP and BOW is also statistically significant, using a Wilcoxon Ranked-Sign test, the difference between BOW and RR is not significant.

We grouped the cases by outcome in order to find out whether there is a difference in performance between cases won by plaintiff and cases won by defendant. A good prediction method, we would expect, has about equal performance for cases won by either side. A method that always predicts the majority class may have high accuracy and coverage, but would be of relatively little use for a practical application. It would



$\mathrm{F}_{\mathrm{pred}}$-Measure	**ProP**	**RR**	**BOW**
Overall	0.703	0.600	0.585
Plaintiff	0. 689	0.678	0. 645
Defendant	0. 724	0.440	0. 459

Fig. 7. $\mathrm{F}_{\mathrm{pred}}$-measure for the classification by SMILE+IBP as a function of case outcome

correspond to an attorney who always advises the plaintiff to sue, and the defendant to settle, irrespective of the case facts.

Comparing prediction performance for the three representations in our experiment, shown in Fig. 7, we found that only ProP satisfies this requirement. Its F-measure is about the same for the cases won by plaintiff and those won by defendant. RR and BOW, on the other hand, have very good performance for the majority class, cases won by plaintiff, but do poorly for the minority class.

To summarize, ProP has significantly better performance than RR and BOW in Experiment II. In addition, ProP is preferable because its performance does not depend on which side won. ProPs are the better representations in both respects; thus, Experiment II supports Hypothesis II.

5 Discussion

The performance of SMILE and SMILE+IBP leaves considerable room for improvement. The average F-measure in Experiment I is below 0.3, which is clearly not sufficient for a practical application. These results raise the question whether the predictions based on automatic Factor assignments have any utility.

In Experiment II, we therefore compared the outputs of SMILE+IBP to an informed baseline. For purposes of the baseline, we assume that it knows the probability of plaintiff and defendant winning. The baseline flips a biased coin, which predicts "plaintiff wins" with the prior probability for plaintiff. This baseline is preferable to the frequently used prediction of the majority class because it takes additional knowledge about the underlying class distribution into account, and because it will have roughly equal performance for cases won by either side, as required above.

SMILE+IBP with Timbl/ProP performs better than this informed baseline: It has an F-measure of 0.70, the baseline's is 0.66. This difference is statistically significant. In interpreting these numbers, one should note that the F-measure is a derived value and does not correspond to any observable features. In particular, we cannot conclude from the difference of 0.04 that SMILE+IBP is "merely 4% more accurate than the baseline." In fact, SMILE+IBP has more than 15% higher accuracy for those cases where it makes a prediction.

We also tested whether one could make equally accurate predictions as SMILE+IBP directly from the case texts, without going through the Factor representation. In a leave-one-out classification experiment with Timbl/ProP, we treated cases won by defendant as positive examples, cases won by plaintiff as negative examples. In 85% of the cases, the classifier predicted "plaintiff wins." The prediction that defendant wins was equally likely for cases won by either side. In effect, the classifier learned to predict the majority class. As argued above, these predictions are of no practical use. We conclude from this experiment that predicting the case outcome directly from the texts is not possible, and that representing cases in terms of Factors in SMILE+IBP is necessary.

In a more informal analysis, IBP's output also provides evidence that SMILE+IBP can generate useful reasoning, despite the many incorrect decisions it made. Shown in Fig. 1, CATO's "gold standard" representation of *National Rejectors* has Factors F7, Brought-Tools (p), F10, Disclosure-to-Outsiders (d), F15, Unique-Product (p), F16, Info-Reverse-Engineerable (d), F18, Identical-Products (p), F19, No-Security-Measures (d), and F27, Disclosure-In-Public-Forum (d).

According to a legal publisher's succinct summary of the court's reasoning, "evidence established that there were no actual trade secrets with respect to plaintiff's [products] and that, although individual defendants, former employees of plaintiff, had improperly used plaintiff's materials and drawings in production of products to compete with plaintiff's products, where plaintiff had not considered information regarding its products to be trade secrets, no warning had been given against use of information." The corresponding issues in IBP's domain model are whether security measures had been taken (Security-Measures), whether the information was valuable (Info-Valuable), and whether defendants had used the information (Info-Used).

As shown in Fig. 5, IBP's analysis of the manually represented case finds all these issues. It correctly reasons that plaintiff is favored for Info-Used, and that defendant is favored for the issue Security-Measures. However, for Info-Valuable, which has Factors for both sides, IBP cannot find sufficient evidence to conclude which side is favored and abstains on that issue. Overall, this analysis matches the court's opinion and leads to a correct prediction.

SMILE+IBP's automatic analysis identifies the same issues, but does not correspond equally well to the court's reasoning because of incorrect Factor assignments. Fig. 2 shows that SMILE found two extra Factors, F6, Security-Measures (p), and F25, Info-Reverse-Engineered (d). It also missed two Factors, F15, Unique-Product (p), and, especially, F27, Disclosure-In-Public-Forum (d), as discussed in Section 4.2. IBP's analysis of *National Rejectors* correctly identifies the issue Info-Used, and comes to the correct conclusion that the plaintiff was favored on the issue. Related to issue Info-Valuable, SMILE assigned only Factor F16, Info-Reverse-Engineered (d). This Factor tends to give inconclusive evidence on this issue and, therefore, is called a weak Factor. IBP takes a conservative approach in this situation; if it finds only a weak Factor, it not does include the issue in its prediction. SMILE also finds the issue Security-Measures. Because of the incorrectly assigned Factor F6, however, IBP cannot resolve the conflicting evidence, which includes Factors F6 and F19; it abstains for the issue. Based on this analysis of the issues, SMILE+IBP abstains for *National Rejectors*.

In processing the example, SMILE+IBP trips over an inconsistency between CATO's representation and SMILE's mark-up conventions. As noted, cases may have textual evidence for F6 as well as for F19 in SMILE+IBP because SMILE can assign both Factors, F6 and F19, to a case. On the other hand, IBP was developed following CATO's conventions that F6 and F19 are mutually exclusive. As a practical matter, it is difficult to implement a principled strategy for SMILE's choosing between F6 and F19 without deeper reasoning and an even more informative knowledge representation. In order to maintain IBP's accuracy and reliability, we do not attempt to resolve the conflict heuristically and let the program abstain on the issue. In a real-world application, a human could easily be alerted if SMILE assigned both F6 and F19 to a case and could determine manually which Factor should apply. Sometimes, indexing may be best handled by a human. Thus, this example raises a more general question for TCBR systems, whether and how best to keep a human in the loop.

6 Summary and Conclusions

This paper introduced SMILE+IBP, a system that integrates methods for assigning abstract indexing concepts to text cases with an interpretive CBR system for argumentation and prediction. The resulting system can carry out reasoning from text cases that goes beyond text retrieval. The goal of our investigation was to identify a good representation for indexing text cases. The experiments showed that both adding background knowledge to replace names and individual entities by their role for the case and using NLP to generate more powerful features, called Propositional Patterns, leads to performance improvements. While our experiments indicate that adding NLP is beneficial, they also pointed to some limitations. Especially for our complex and hard-to-parse texts, NLP remains a bottleneck to which many errors can be attributed, even though we had a robust, high-performance parser. Further, our experiments suggest that ProPs are most beneficial for Factors that correspond to relatively complex fact situations. On the other hand, for simpler fact patterns, like our F15, a representation like RR, that does not rely on NLP, may be suitable.

Among three different learning algorithms, NN had the best performance. Our data are very skewed and sparse, which makes it difficult to find patterns or generalize from the examples. Under these circumstances, NN did a better job identifying the most relevant examples, especially for the harder-to-find concepts.

SMILE+IBP has not reached the level of performance that would be required by attorneys. It is a step in that direction, however, in that it integrates indexing and reasoning with text cases. Despite all SMILE's limitations, the program does significantly better than an informed baseline. Moreover, as illustrated in the *National Rejectors* example, IBP is fairly robust. IBP+SMILE's analysis of the automatically indexed case is reasonable and identifies the major issues. Due to errors by SMILE, IBP abstains, indicating that human intervention may be required. Indexing text cases is a hard problem; automatic indexing will always be subject to certain limits and a human may need to tackle the harder problems of text interpretation.

Acknowledgements

This research has been supported by the National Science Foundation under award IDM-9987869. We would like to thank Ellen Riloff for giving us access to AutoSlog/Sundance, which have been a relevant resource and inspiration for this project.

References

Aleven, V. 2003. Using Background Knowledge in Case-Based Legal Reasoning: A Computational Model and an Intelligent Learning Environment. *Artificial Intelligence* 150(1-2):183–237.

Ashley, K., and Brüninghaus, S. 2003. A Predictive Role for Intermediate Legal Concepts. In *Proc. 16th Annual Conference on Legal Knowledge and Information Systems*.

Ashley, K. 1990. *Modeling Legal Argument, Reasoning with Cases and Hypotheticals*. MIT-Press.

Brüninghaus, S., and Ashley, K. 1999. Bootstrapping Case Base Development with Annotated Case Summmaries. In *Proc. 3rd International Conference on Case-Based Reasoning*.

Brüninghaus, S., and Ashley, K. D. 2001. The Role of Information Extraction for Textual CBR. In *Proc. 4th International Conference on Case-Based Reasoning*.

Brüninghaus, S., and Ashley, K. D. 2003. Combining Case-Based and Model-Based Reasoning for Predicting the Outcome of Legal Cases. In *Proc. 5th International Conference on Case-Based Reasoning*.

Burke, R.; Hammond, K.; Kulyukin, V.; Lytinen, S.; Tomuro, N.; and Schonberg, S. 1997. Question-Answering from Frequently-Asked Question Files: Experiences with the FAQ-Finder System. *AI Magazine* 18(1):57–66.

Burke, R. 1998. Defining the Opportunities for Textual CBR. In *Proc. AAAI-98 Workshop on Textual Case-Based Reasoning*.

Cardie, C., and Howe, N. 1997. Improving Minority Class Prediction Using Case-Specific Feature Weights. In *Proc. 14th International Conference on Machine Learning*.

Cohen, P. 1995. *Empirical Methods for Artificial Intelligence*. MIT-Press.

Cunningham, C.; Weber, R.; Proctor, J. M.; Fowler, C.; and Murphy, M. 2004. Investigating Graphs in Textual Case-Based Reasoning. In *Proc. 7th European Conference on Case-Based Reasoning*.

Daelemans, W.; Zavrel, J.; van der Sloot, K.; and van den Bosch, A. 2004. TiMBL: Tilburg Memory Based Learner, version 5.02. http://ilk.kub.nl/software.html.

Daniels, J., and Rissland, E. 1997. Finding Legally Relevant Passages in Case Opinions. In *Proc. 6th International Conference on Artificial Intelligence and Law*.

Dietterich, T. 1996. Statistical Tests for Comparing Supervised Classification Learning Algorithms. Oregon State University Technical Report.

Gupta, K., and Aha, D. W. 2004. Towards Acquiring Case Indexing Taxonomies from Text. In *Proc. 6th International Florida Artificial Intelligence Research Society Conference*.

Lenz, M. 1999. *Case Retreival Nets as a Model for Building Flexible Information Systems*. Ph.D. Dissertation, Humboldt University, Berlin, Germany.

McCallum, A. K. 2004. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ ∼mccallum/bow.

Quinlan, R. 2004. C4.5 Release 8. http://www.rulequest.com/Personal/.

Riloff, E. 2003. From Manual Knowledge Engineering to Bootstrapping: Progress in Information Extraction and NLP. Invited Talk at the Fifth International Conference on Case-Based Reasoning (ICCBR-03), http://www.iccbr.org/iccbr03/invited.html.

Rose, D. 1994. *A Symbolic and Connectionist Approach to Legal Information Retrieval*. Hillsdale, NJ: Lawrence Earlbaum Publishers.

Salzberg, S. 1997. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery* 1(3):317–328.