Toward Automated Definition Acquisition From Operations Law

Yi Chang, *Member, IEEE*, Jana Diesner, *Member, IEEE*, and Kathleen M. Carley, *Member, IEEE*

***Abstract*—Definition acquisition is a necessary step in building an artificial cognitive assistant that helps military personnel to gain fast and precise understanding of the various terms and procedures defined in applicable legal documents. We approach the task of identifying definitional sentences from operations law documents by formalizing this task as a sentence-classification task and solving it by using machine-learning methods. This paper reports on a series of empirical experiments in that we evaluate and compare the performance of learning algorithms in terms of label-prediction accuracy. Using supervised techniques results in an F1 score of 95.93% and a 96.72% recall rate. However, for real-world applications, it would be too costly and unrealistic to ask personnel involved in military operations to label substantial amounts of data in order to build a new classifier for different types or genres of text data. Therefore, we propose and implement a semisupervised (SS) solution that trades off prediction accuracy to label efficiency. Our SS approach achieves a 90.47% F1 score and 93.44% recall rate by using only eight sentences labeled by a human expert.**

***Index Terms*—Definitional-sentence classification, semisupervised (SS) learning, supervised learning.**

I. Introduction

TYPICALLY, the conduct of military operations is governed by laws, rules, and agreements. For the U.S. military, the regulations applicable in the case of military operations and contingencies on U.S. territory and abroad are defined in the *operations law* [24]. The operations law is a collection of domestic, foreign, and international regulations. It includes key legislations regarding conflicts, such as *the law of armed conflict* (LOAC), which is also referred to as *the law of war*, which is a part of international law that regulates the conduct of armed hostilities [27]. For any specific operation, which can range from peacetime missions to armed conflicts, military authorities issue standing *Rules of Engagement* (RoE). The RoE "delineate the circumstances and limitations under which the United States forces will initiate and/or continue combat engagement with other forces encountered" [25]. Different nations and transnational organizations, such as the North Atlantic Treaty Organization (NATO), often have their own RoE.

In the case of deployment abroad and involvement in international conflicts, any player involved, including commanders, judge advocates, and peace-keeping forces, are required to consider international laws in addition to national regulations. As part of this process, each player has to quickly gain a correct understanding of the terms and procedures defined in any operations law that apply to the current situation. The relevance of a regulation can vary from operation to operation. These requirements can lead to a large collection of a different set of documents to be taken into account for every new operation by groups and units of any size within a short amount of time. The broader goal with the research presented herein is to support this knowledge-acquisition process through reliable technologies [19]. In this paper, our specific goal is to develop a labeling efficient and predictably accurate system that detects and extracts the definitions of concepts from a specific convention, namely, the *United Nations Convention on the Law of the Sea* (UNCLOS). UNCLOS regulates various aspects pertaining to the seas and the oceans, such as territorial, economic, and ecological questions, including fishery and piracy [26]. UNCLOS was passed after nine years of negotiation as an international customary law in 1982 and is currently ratified by 157 states. Customary laws codify rules that are legally binding to those who ratify them [1]. Even though the U.S. did not sign the full UNCLOS treaty, the RoE for a U.S. Navy mission might order the respective team to comply with UNCLOS.

Machine-learning methods are powerful techniques to solve information-extraction tasks, including definition extraction. Therefore, we approach the definitional sentence-classification problem by using different learning methods. For real-world applications, it is inefficient and impractical to ask command authorities or legal experts to label large amounts of data, which would be necessary to train supervised models. To minimize the human-labeling effort, we propose a semisupervised (SS) learning solution that balances prediction accuracy and labeling efficiency. How can this approach serve humans in their decision making efforts? On the long run, we envision our solution to integrate with an artificial cognitive assistant that quickly and predictably accurately extracts definitions from new legal documents based on a few rounds of interactions with a human subject-matter expert through an interactive interface as follows: In each round, the assistant presents the human expert with a few sentences and asks her to select an appropriate, predefined label per sentence, such as "definitional" or "nondefinitional." After a few rounds, the assistant uses the expert's choices

Manuscript received February 20, 2010; revised May 19, 2010, August 21, 2010, and November 11, 2010; accepted January 8, 2011. Date of publication March 10, 2011; date of current version February 17, 2012. This work was supported by the Army Research Institute under Grant W91WAW07C0063, by the Office of Naval Research under Rules of Engagement N000140610104, N000141010610, and MURI N000140811186, and by the Center for Computational Analysis of Social and Organizational Systems, Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Institute, the Department of Defense, or the U.S. government. This paper was recommended by Associate Editor S. H. Rubin.

The authors are with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: changyi@cs.cmu.edu; diesner@cs.cmu.edu; kathleen.carley@cs.cmu.edu).

Digital Object Identifier 10.1109/TSMCC.2011.2110643

1094-6977/$26.00 © 2011 IEEE

to train a classifier. In summary, the ultimate goal here is to develop an efficient and effective solution that generalizes well to previously unseen and potentially large amounts of operations law documents so that people with a need for definitions of terms are supported.

The contributions of the research presented herein include the following:

1) identification of surface patterns as features from natural language text data that are suitable inputs to a supervised machine-learning system which achieves high accuracy in detecting and classifying definitional sentences from operations laws;
2) proposal and evaluation of an SS learning solution that minimizes human-labeling efforts for military applications while achieving high accuracy.

The remainder of this document provides a concise background section on the more general definition-extraction task. We continue with reporting on how we utilize state-of-the-art machine-learning techniques to solve the definition-extraction task for the domain of legal text documents and on the performance and evaluation of the proposed solution.

II. Problem Formalization

Definition acquisition from legal documents is related to question answering (QA), which is an information-retrieval task [20]. In QA, definitions are formalized as those sentences that contain the most-descriptive information; for example, answering questions like "What is XX?" or "Who is YY?" requires the extraction of multiple answers from multiple documents and the combination of the extracted answers into a single unified answer. Definitional QA problem has been solved via surface-pattern-based methods [1], [5], [9], [18], [22], and have been formalized as a ranking problem [7], [21].

Table I contains several examples of definitional sentences from UNCLOS. From reading this corpus, we have identified characteristics that are specific to this body of law. These characteristics differ from traditional definitional QA based on news articles.

1) *Specialty:* A common concept can be assigned to a specific meaning in the context of law. For example, in UNCLOS, the concept of "*area*" is defined as *the seabed and ocean floor and subsoil thereof beyond the limits of national jurisdiction*.
2) *Inheritance:* A specific concept can inherit part of its definition from a more generic parent concept. For example, in UNCLOS, "*pioneer area*" is defined as *an area allocated by the Commission to a pioneer investor for pioneer activities pursuant to this resolution*. A pioneer area is a specific area such that the definition of the concept "*area*," as provided above, also applies.
3) *Comprehensiveness:* A sentence defining a single concept can be very long; occasionally spanning more than a single page. Definitions, furthermore, include specifications of what something is and what it is not. For an example, see the UNCLOS definition of "*dumping*" in Table I.

In this paper, we focus on identifying explicit definitions from military operations laws. Besides explicit definitional sentences, UNCLOS also contains sentences with implicit definitional description, such as the examples shown in Table II. These sentences are not considered by the identification technology described herein.

TABLE I
Examples of Definitional Sentences in UNCLOS

Definition Concept	Definition Sentence
area	*``Area" means the seabed and ocean floor and subsoil thereof, beyond the limits of national jurisdiction.*
pioneer area	*``Pioneer area" means an area allocated by the Commission to a pioneer investor for pioneer activities pursuant to this resolution.*
bay	*For the purposes of this Convention, a bay is a well-marked indentation whose penetration is in such proportion to the width of its mouth as to contain land-locked waters and constitute more than a mere curvature of the coast.*
natural resources	*The natural resources referred to in this Part consist of the mineral and other non-living resources of the seabed and subsoil together with living organisms belonging to sedentary species, that is to say, organisms which, at the harvestable stage, either are immobile on or under the seabed or are unable to move except in constant physical contact with the seabed or the subsoil.*
dumping	*"dumping" means:* *(i) any deliberate disposal of wastes or other matter from vessels, aircraft, platforms or other man-made structures at sea;* *(ii) any deliberate disposal of vessels, aircraft, platforms or other man-made structures at sea;* *"dumping" does not include:* *(i) the disposal of wastes or other matter incidental to, or derived from the normal operations of vessels, aircraft, platforms or other man-made structures at sea and their equipment, other than wastes or other matter transported by or to vessels, aircraft, platforms or other man-made structures at sea, operating for the purpose of disposal of such matter or derived from the treatment of such wastes or other matter on such vessels, aircraft, platforms or structures;* *(ii) placement of matter for a purpose other than the mere disposal thereof, provided that such placement is not contrary to the aims of this Convention.*

TABLE II
Examples of Nondefinitional Sentences in UNCLOS

A coastal State which is a member of or has a bilateral agreement with an international organization, and in whose exclusive economic zone or on whose continental shelf that organization wants to carry out a marine scientific research project, directly or under its auspices, shall be deemed to have authorized the project to be carried out in conformity with the agreed specifications if that State approved the detailed project when the decision was made by the organization for the undertaking of the project, or is willing to participate in it, and has not expressed any objection within four months of notification of the project by the organization to the coastal State.
A State Party which is in arrears in the payment of its financial contributions to the Authority shall have no vote if the amount of its arrears equals or exceeds the amount of the contributions due from it for the preceding two full years.

We separate the definitional knowledge-acquisition process into two tasks: definition detection and defined-concept extraction. Definition detection means to identify sentences that define concepts from natural-language text. Defined-concept extraction means to locate and extract the actual concept that is defined in a definitional sentence.

Defined-concept extraction is a relatively simple and straightforward task: Given a definitional sentence, several heuristics and rules can be applied to successfully extract the defined concepts from definitional sentences. Following is an example for these rules: Pick the quoted named entity right next to either one of the following term “means/considers/refers to.”

Definition detection from a large operations law corpus is a more challenging problem. In this paper, we formulate the definition-detection task as a sentence-level-classification problem. The evaluation of the identified sentences is straightforward: A false positive is a sentence that does not define a concept, although it might contain some relevant information about a concept. A false negative is a definitional sentence that we failed to identify as such. In the given application domain, the latter type of error seems more severe as it may result in disregarding factual knowledge and the noncompliance with RoE. Therefore, for definition detection from operation-law documents, the recall rate is a more crucial performance measure than the precision rate. These measures are explained in Section IV-C.

III. Related Work

The majority of related work stems from definitional QA. Prior work in definitional QA can be categorized into three families of methodological approaches, which are often combined into multimethod approaches for practical applications: surface-pattern-based methods, ranking-based methods, and Internet-data-driven methods.

A. Surface-Pattern-Based Methods

Blair-Goldensohn *et al.* [1] presented a definitional QA system that combines knowledge-based methods with statistical methods. Cui *et al.* [9] explored the usage of probabilistic lexical–syntactic pattern matching for definitional QA. Ravichandran and Hovy [18] developed a method to learn surface patterns via bootstrapping, which is an SS learning method. Overall, surface-pattern-based methods are a simple and straightforward approach, but they cannot achieve the accuracy rates comparable with those from alternative approaches [6], [22].

B. Ranking-Based Methods

Chen *et al.* [7] proposed a reranking method for answers that is novel in its language model, which considers dependencies between words. Their method outperforms prior bag-of-words-based approaches. Xu *et al.* [21] ranked retrieved definitional excerpts according to the expert’s likelihood of being good definitions. Their experiments show that ranking outperform classification and ordinal regression. Ranking-based methods are highly dependent on the modeling algorithms and do not leverage other sources of information.

C. Internet-Data-Driven Methods

Cui *et al.* [8] demonstrated an approach that applies soft-matching patterns to text data from the Internet in order to identify definitional sentences. Hildebrandt *et al.* [12] proposed a multimethod approach to QA that combines an offline database constructed from surface patterns with searching through an online dictionary plus using an off-the-shelf retriever for documents from the Internet. Lampouras and Androutsopoulos [14] generated training examples for supervised learning from online encyclopedias and text data, and based on that train, a model that is then used to search the web for definitions of concept that have yet to be covered by the online encyclopedias. Given the characteristics of operations law, we cannot leverage information from the Internet for the given definition detection and extraction task.

IV. Definition Detection With Supervised-Learning Methods

As prior work has shown, one promising strategy to tackle the definitional sentence-classification problem is to use supervised machine-learning methods that share the same data representation. For this purpose, each sentence is represented as a feature vector, where each feature provides some signal that increases or decreases the overall likelihood of a sentence to be definitional.

Our supervised procedure consists of the following steps: Given an operation-law document, we first split the document into sentences. Then, we predict for every sentence whether it is definitional or not. If it is a definitional sentence, the defined concept is extracted from it.

Sentence splitting and defined-concept extraction can be solved with high accuracy by using well-defined heuristics and rules. In this paper, we focus on the accurate and efficient prediction of definitional sentences. We apply generative learning method (i.e., Naïve Bayes) and discriminative learning methods (i.e., support-vector machines (SVMs) and gradient-boosting decision tree) to solve the definition-detection problem.

In contrast with traditional definitional QA tasks, definition detection and extraction from operations law requires high precision. Furthermore, these documents are lengthier and more complex than documents on the web typically are. To the best of our knowledge, there is no prior work on definition detection and extraction from operations-law text data.

A. Supervised Learning Algorithms

1) Naïve Bayes Model: Naïve Bayes model [15] is a simple and effective generative learning method for classification tasks, and we include it as our baseline. The basic idea is the Bayes theorem is

$$P(y|x) = \frac{P(y) \times P(x|y)}{P(x)}. \tag{1}$$

Each sentence is represented as a vector of Nfeatures of the observed language input $x = (v_1, v_2, \ldots, v_N)$.

To simplify the computation, the Naïve Bayes model assumes that all attribute values v_i are independent, i.e., for $i \neq j$, v_i and v_j are conditional independent of given x. Therefore, (1) could be simplified to

$$P(y|x) = \frac{P(y) \times \prod_{i=1}^{N} P(v_i|y)}{P(x)}. \tag{2}$$

Based on (2), a Naïve Bayes model can be constructed according to the rule of maximum *a posteriori* (MAP). The corresponding classifier is defined as follows:

$$y^* = \arg\max_{y \in Y} \{P(y) \times \prod\nolimits_{i=1}^{N} P(v_i|y)\}. \tag{3}$$

To avoid $P(v_i|y)$ to equal zero, we use the Laplace smoothing method [7] for our experiments.

2) Support-Vector Machine: SVM is the most widely used method for classification. SVM searches for a hyperplane that separates a set of positive and negative training examples [3], which correspond to definitional and nondefinitional sentences in our experiments. The hyperplane is defined as $w^T x + b = 0$, where $w \in R^d$ is a vector orthogonal to the hyperplane, $|b|/||w||$ is the perpendicular distance from the hyperplane to the origin, and $||w||$ is the Euclidean norm of w. The decision function is the hyperplane classifier given by

$$F(x) = \mathrm{sign}(w^T x + b). \tag{4}$$

The hyperplane is subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, N, \qquad \xi > 0 \tag{5}$$

where $x_i \in R^d$ is a training example with d dimensions of features, and $y_i \in \{+1, -1\}$ denotes the label of the feature vector x_i. In this formula, ξ is a positive slack variable, and sum of ξ_i means the upper bound of training errors. The margin is defined by the distance between two parallel hyperplanes $w^T x + b = 1$ and $w^T x + b = -1$. Therefore, the SVM training process can be defined as an optimize problem as follows:

$$\text{Minimize}\left(\frac{1}{2} w^T w + \gamma \sum_i \xi_i\right) \tag{6}$$

where γ is the regularization parameter, which is usually empirically selected to reduce the testing errors. Equation (6) is subjected to (5).

A linear SVM classifies linearly separable data points on a hyperplane. If the data points are not linearly separable, the basic SVM can be extended by nonlinear kernels such that high-dimensional hyperplanes can be processed as

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^p \tag{7}$$

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \quad \text{for } \gamma > 0 \tag{8}$$

$$K(x_i, x_j) = \tanh(k x_i \cdot x_j - \delta), \quad \text{for } \delta > 0. \tag{9}$$

To implement SVM, we use the SVM-light package [13]. We use a linear kernel function as the baseline, a polynomial kernel function [see (7)] with $p$degree, a radial-basis kernel function [see (8)], and a sigmoid kernel function [see (9)].

However, nonlinear SVM is heavily dependent on the data distribution. If the data do not fit well on those predefined SVM nonlinear kernels on high dimension, nonlinear SVM will not perform well. Therefore, we also choose gradient-boosting tree (GBT), which is a gradient-boosting nonlinear model, but do not rely on data distribution.

3) Gradient-Boosting Tree: The basic idea with boosting is to iteratively decrease the error, which is also called loss function, with a weak learner. GBTs use decision trees as weak learners [10] and iteratively fit an additive model in order to minimize the loss function $L(y_i, f_T(x + i))$ as

$$f_t(x) = \mathrm{TR}_t(x; \theta_0) + \lambda \sum_{t=1}^{T} \beta_t \mathrm{TR}_t(x; \theta_t) \tag{10}$$

where $\mathrm{TR}_t(x; \theta_t)$ is a decision tree at iteration t, which is weighted by parameter β_t, θ_t is the parameter in the decision trees, and λ is the learning rate. At iteration t, tree $\mathrm{TR}_t(x; \theta_t)$ is generated to fit the negative gradient of least-square errors as

$$\hat{\theta} = \arg\min_{\beta} \sum_{i}^{N} (-G_{it} - \beta_t \mathrm{TR}_t(x); \theta)^2 \tag{11}$$

where G_{it} is the gradient over the current prediction function given by

$$G_{it} = [\partial L(y_i, f(x_i))/\partial f(x_i)]_{f = f_{t-1}}. \tag{12}$$

The optimal weight of trees β_t is computed as

$$\beta_t = \arg\min_{\beta} \sum_{i}^{N} L(y_i, f_{t-1}(x_i) + \beta \mathrm{TR}_t(x_i, \theta). \tag{13}$$

If we choose squared errors as the loss function, the gradient is given by $G(x_i) = -y_i + f(x_i)$.

Compared with nonlinear SVM kernels, GBT is more robust for nonlinear classification tasks since the "kernel trick" in SVM relies on the distribution of the data. If the data do not fit the predefined SVM kernels well on a high-dimensionality hyperplane, GBT outperforms nonlinear SVM [4].

B. Data and Features

We evaluate the performance of different supervised learning algorithms on UNCLOS. First, UNCLOS was split up into 3949 sentences. A total of 807 of those sentences are empty or only contain page-index information. For this project, we established a ground truth to learn from and evaluate against by labeling the remaining 3142 sentence in UNCLOS as being definitional or not. This process was pursued by all the three authors collaboratively until an agreement on every sentence was reached. As a result, we identified 61 definitional sentences and 3081 nondefinitional sentences.

Since the total number of positive examples is small, we need to avoid a high-dimensionality feature space. Furthermore, in order to build a robust solution, all features should generalize well to new and unseen legal documents. To solve this task, we propose and implement a set of contextually sensitive features generated from surface patterns of the text data.

Much of the prior work in definitional QA uses surface patterns for definition extraction. Those surfaces can either be manually generated [5], [20], or be automatically learned [1], [9], [18]. Prior examples for such surface patterns are as follows:

– <Definition>, <Defined Term>.
– <Defined Term> is a <Definition>.
– <Defined Term>, which is <Definition>.

Surface-pattern-based methods are typically reported to result in accuracy rates far below 90%, which is unacceptable for critical missions. While relying on surface patterns alone might not

TABLE III
FEATURES EXPLANATION FOR DEFINITION SENTENCE CLASSIFICATION

ID	Feature Type	Feature Explanation
1	*Count*	*Number of quotation marks in the sentence;*
2	*Count*	*Number of words in the <key-term>;*
3	*Count*	*Number of words in the sentence;*
4	*Binary*	*<key-term> occurs at the beginning of the sentence;*
5	*Binary*	*<key-term> occurs at the end of the sentence;*
6	*Binary*	*<key-term> begins with "the" or "a" or "an";*
7	*Binary*	*<key-term> is initial letter uppercase;*
8	*Binary*	*<key-term> contains pronouns;*
9	*Binary*	*<key-term> is followed by "is a" or "is an" or "is the";*
10	*Binary*	*<key-term> is followed by "which is" or "which are"*
11	*Binary*	*The sentence contains "mean" or "means" or "meant";*
12	*Binary*	*The sentence contains "means of";*
13	*Binary*	*The sentence contains "refer" or "refers" or "referred";*
14	*Binary*	*The sentence contains "consider" or "considers" or "considered";*
15	*Binary*	*<key-term> does not occur in the global <key-term> list;*
16	*Binary*	*<key-term> contains new context word that does not occur in the global context word list;*

lead to satisfactory accuracy, using them as additional features may provide informative signals to the learning mechanism. Therefore, we convert existing surface patterns that have shown to be useful for definition extraction into features for learning algorithms. In order to facilitate feature generation, we assume that every sentence contains exactly one <key-term>, which we assume to be the same as the main term of the sentence. We apply the following heuristic rules to identify one <key-term> per sentence.

1) If a sentence contains one or multiple pairs of quotation marks, the first quoted phrase is picked as the <key-term>.
2) Otherwise, the first noun phrase in the sentence is identified as the <key-term>.

In UNCLOS, the remaining sentences are typically either empty or only contain page index information. The remaining sentences are disregarded for further work.

For a definitional sentence, the defined concept does not necessarily match the <key-term> of the sentence. That is ok since the <key-term> is mainly used for feature generation.

Table III shows the list of the 16 features that we generated for definitional sentence classification. All but the last two features are generated based on surface information from the data. The last two features are generated based on global information from UNCLOS: while all sentences are processed one by one from the beginning to the end, a global <key-term> list and a global <key-term> context list are compiled. We define context words as the three words left and three words right to the <key-term>. If the <key-term> for a given sentence does not yet occur in the global <key-term> list, the value for the pre and last feature is set to 1; otherwise, it is set to 0. If none of the context words for a given sentence occur yet in the global context list, the value for the last feature is set to 1; otherwise, it is set to 0. The basic idea with last two features is that defined concepts are most likely to be embedded in a definitional sentence the first time they occur in the data.

TABLE IV
METRICS DESCRIPTION

	Definitional (label)	Non-Definitional (label)
Definitional (prediction)	True Positive	False Positive
Non-Definitional (prediction)	False Negative	True Negative

TABLE V
EVALUATION OF SUPERVISED METHODS WITH FIVEFOLD CROSS VALIDATION

	Precision (%)	Recall (%)	F1 Score (%)
Naïve Bayes	73.52	78.69	76.02
SVM (Linear)	88.89	88.52	88.70
SVM (Polynomial)	89.70	96.72	93.08
SVM (Radial Basis)	90.91	95.08	92.94
SVM (Sigmoid)	81.42	93.44	87.02
GBT	95.16	96.72	95.93

C. Evaluation

For classification tasks, the commonly used evaluation metrics are precision, recall, and F1 score: Intersecting the predicted label for a sentence with the ground truth for the label of the same sentence results in four categories of classification results, as shown in Table IV—true positive (tp), true negative (tn), false positive (fp), and false negative (fn). Based on that, precision, recall, and the F1 score are computed as follows:

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{14}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{15}$$

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{16}$$

F1 score is also called the harmonic average or harmonic mean of precision and recall. In our experiments, the sum of true positive (tp) and false negative (fn) is 61, and the sum of true negative (tn) and false positive (fp) is 3081. Therefore, the key task with evaluating our findings is to figure out the false positive (fp) and false negative (fn).

D. Experimental Result

Table V shows the accuracy rates resulting from using the supervised methods described earlier. These experiments are based on fivefold cross validation. The results show that the discriminative models outperform the generative models in both precision and recall. Comparing the different kernels of SVM, we observe that the two nonlinear kernels outperform linear kernels. The polynomial kernel performs best, while the sigmoid kernel performs worst on this dataset. Out of all algorithms tested, GBT results in the highest accuracy rates. As explained in Section II, for this project, recall is more important than precision. The numbers in Table V demonstrate that GBT and all nonlinear SVM methods achieve high recall rates. The best result has a total of two false-negative sentences and three false-positive sentences. Our error analysis reveals that in UNCLOS, several definitional sentences do not contain any explicit functional word or patterns that are captured with our feature set. Therefore, the detected errors might be fixable by adjusting the feature set.

In the next section, we explore using an SS approach in order to simulate the intended real-world application. Based on the empirical findings from this section, we will compare our results only with the highest performing algorithms from this section, namely, SVM with polynomial kernel, SVM with radial-basis kernel, and GBT.

V. Definition Detection With Minimum Human-Labeling Efforts

In the context of real-world, real-time military operations, it would be prohibitively costly and time-consuming for the involved personnel to label training data. An appropriate alternative to the supervised approach is SS learning, which aims to minimize the human-labeling efforts without severely limiting the prediction accuracy. For this paper, we investigate the usage of SS learning to extract definitions from UNCLOS, thereby taking into account the constraints of the given domain and aiming to balance labeling efficiency and prediction accuracy.

SS learning is a set of statistical machine-learning techniques that exploit a small amount of labeled data and a large amount of unlabeled data for training [2], [11]. There are two key questions for SS learning: First, how to select the best candidates to be labeled? Second, how to best combine labeled and unlabeled data? For our experiments, we assume all sentences to be unlabeled in the beginning. We then label a small amount of sentences in order to represent the input needed from a human.

A. Semisupervised Learning Solution

In this section, we propose an SS learning algorithm that efficiently and effectively combines labeled data with unlabeled data while minimizing human-labeling efforts. The algorithm requires several iterations: At each step, one or multiple labeled sentences are added to the set of positive examples. After that, the learned model is updated. When the most-recent model is applied to the set of unlabeled data, the most-ambiguous unlabeled data points near the decision boundary are identified and presented to a human expert for judgment. The rationale for this decision is that labeling these sentences provides the most amount of information to the weak learner. The selection of sentences to be labeled by a human can be further improved by excluding sentences that are very likely to be negative examples, i.e., nondefinitional sentences. To be consistent with the overall methodology presented herein, we use surface patterns and heuristics for this step. For example, in UNCLOS, sentences containing “should” are likely to be nondefinitional.

Besides relying on human judgment, the algorithm also decides automatically about the category for a sentence by computing a confidence value for the predicted label for a new example. If this value is higher than a preset threshold, the sentence is classified. As shown in Algorithm 1, our SS algorithm consists of three steps.

```
Algorithm 1:
1. Set positive set $L_p$ and negative set $L_n$ as empty.
   Select $r$ candidate positive examples $e$ according to surface patterns and request label from human expert. For each candidate:
      if $e$ is labeled as positive, put it into $L_p$, then randomly select $\alpha$ negative examples and put them into $L_n$;
      if $e$ is labeled as negative, put it into $L_n$;
   Train an initialization classifier $M$ with $L_p$ and $L_n$;
2. $i \Leftarrow r$
   repeat
      $i \Leftarrow i+1$
      Test model M on the unknown set, then choose the most ambiguous positive example $e$ and request label for them from human expert;
      if $e$ is labeled as positive, add it to $L_p$, then randomly select $\alpha$ negative examples put to $L_n$;
      if it is labeled as negative, put it into $L_n$;
      Update model $M$ according to $L_p$ and $L_n$;
   Until $i = r + l$
3. repeat
      Test model M on the unknown set, and choose the most confident example;
      If it is predicted as positive, and confidence is high, add it into $L_p$;
      If it is predicted as negative, and confidence is high, add it into $L_n$;
      Update model $M$ according to $L_p$ and $L_n$;
   Until $i = N$
```

1) *Initialization:* We draw a random sample from the pool of candidates that are highly likely to be positive examples according to the established heuristics and ask a human expert to label these examples. We then add a few negative examples as input to the learner. We use the set of positive examples identified by the human plus the set of automatically detected negative examples as the training data to train a first classifier model M.

2) *Active learning to find the best candidates for human labeling:* At each iteration, train a classifier model $M$by using the existing training data. Test M on the test data. Select the most-ambiguous candidates predicted to be positive examples, and collect human feedback on them. This

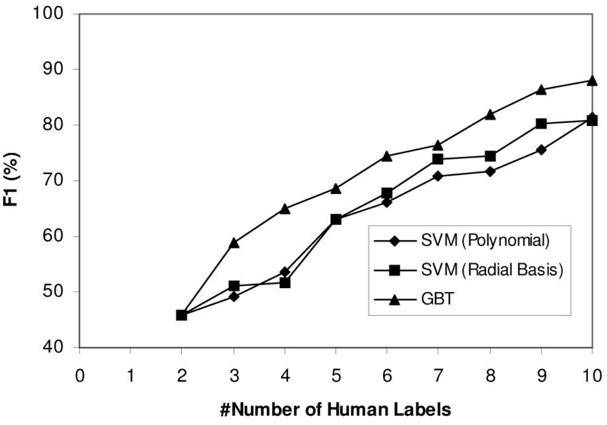

Fig. 1. F1 score comparison with different number of human labels.

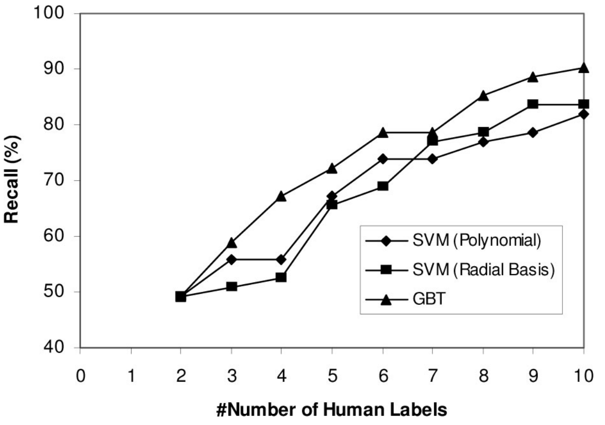

Fig. 2. Recall comparison with different number of human labels.

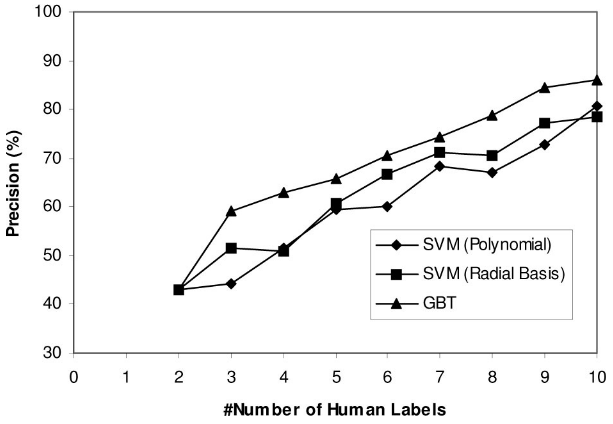

Fig. 3. Precision comparison with different number of human labels.

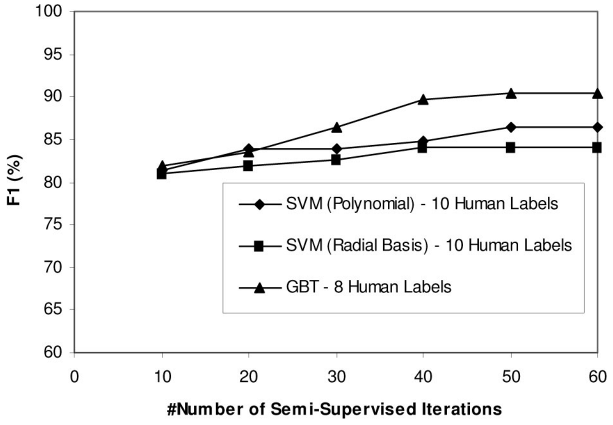

Fig. 4. F1 score with different number of SS iterations.

is a sequential process, i.e., the hand-labeled data from the current iteration will be used to train M in the next iteration.

3) *Automatic label prediction:* In each iteration, train a classifier model M with the existing training data, and test it on the unknown dataset, and then select the examples predicted with high confidence into the corresponding training set. This is also a sequential process, i.e., the data predicted and added to the training set in the current iteration will be used to train M in the next iteration.

Furthermore, for the first and second steps, we establish the following rule to balance the total amount of positive and negative examples: In each iteration, the same number α of sentences that the human labels as positive is also added to the negative set. In order to minimize human-labeling efforts, we need to keep $r + l$ as small as possible.

A couple of previous works target to combine active learning and SS learning together, which is equal to combine steps 2 and 3 into one step. Unfortunately, the combined method is implemented with an EM algorithm [16], or cotraining algorithm [17], or Gaussian random fields [23], and all of them need multiple iterations to converge, which is time-consuming and not suitable for real-time application. In addition, starting off with such a mixture of labeled and unlabeled examples increases the chance of a semantic drift, which means the exponentially increasing gravitation of the prediction model into an erroneous direction.

B. Experimental Results

For our experiments, we set the negative/positive ratio α to 5, the number of preliminary positive example r to 2, and the total number of human labels to provide l to be smaller or equal to 10.

The experiment designed in this section is an offline simulation experiment. Using UNCLOS as ground-truth data, we simulate the SS learning method. In order to do this, we assume no labels to be available in the beginning, and then run the algorithms with different parameter values to evaluate accuracy and the total number of labels required to train the classifier model.

Figs. 1–3 show the F1 score, recall, and precision for $r + l$. The results suggest that as the number of labels provided by humans increases, the accuracy of classification rises. With ten human labels, SVM with a polynomial kernel has an F1 score of 81.30% and 81.97% recall rate. With ten human labels, SVM with a radial-basis kernel has an F1 score of 80.95% and 83.61% recall. With eight human labels, GBT has an F1 score of 88.00% and 90.16% recall.

Figs. 4–6 show the development of accuracy for the three algorithms mentioned in the previous paragraph without requesting human labels beyond the first round. According to these

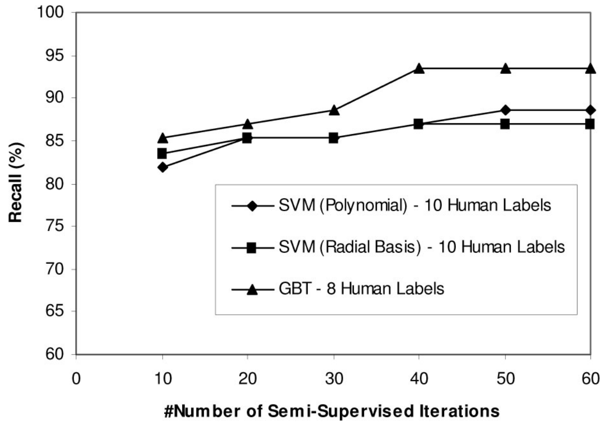

Fig. 5. Recall comparison with different number of SS iterations.

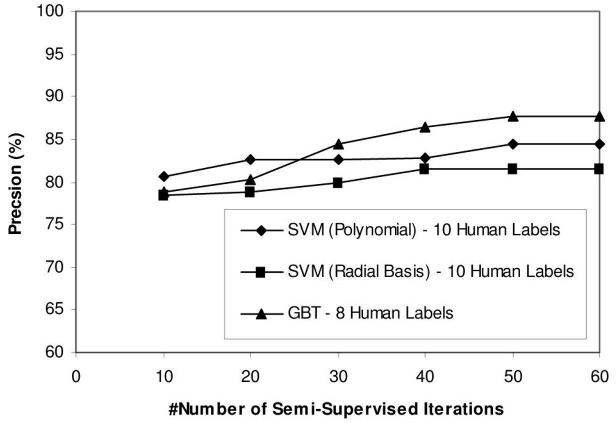

Fig. 6. Precision comparison with different number of SS iterations.

figures, the improvements are marginal for SVM with a radial-basis kernel but are meaningful for the other two algorithms. For GBT, the F1 score increases from 81.90% at the eighth iteration to 90.47% at the 60th iteration and corresponding recall increase from 85.26% to 93.44%; for SVM with polynomial kernel, the F1 score increases from 81.30% at the tenth iteration to 87.09% at the 60th iteration; however, for SVM with radial-basis kernel, the F1 score only increases from 80.95% at the tenth iteration to 84.13% at the 60th iteration. Overall, GBT outperforms the nonlinear SVM for both, i.e., predication accuracy and reducing human efforts.

The error analysis on the unsupervised results reveal that the ninth feature (i.e., sentences containing “is a” or “is the”) is difficult to be selected and classified because this surface pattern is also common in nondefinitional sentences. Without labeling the majority of the positive examples, the model could not predict respective sentences accurately. For this feature, the number false negative is always greater than zero. However, due to the limited amount of human labels, the majority of positive examples for this feature could not be labeled. This issue is the major cause for the accuracy gap between the best-supervised learning model and the best- SS model. We also found that as long as the total number of labels is small, sentences with “which,” “is a,” and “is the” are the most-difficult ones to correctly classify.

TABLE VI
ACCURACY COMPARISON WITH TOTAL AMOUNT OF HUMAN JUDGMENTS

	# Judgments	Precision (%)	Recall (%)	F1 Score (%)
¯Random 10%	312	85.71	77.78	81.55
Random 20%	624	82.35	79.25	80.77
Random 30%	936	83.33	86.96	85.11
Random 40%	1248	87.18	85.0	86.08
Random 50%	1560	93.75	88.25	90.92
Random 60%	1872	88.46	92.0	90.20
Random 70%	2184	94.73	90.0	92.30
Random 80%	2496	95.16	96.72	95.93
Random 90%	2808	95.16	96.72	95.93
GBT+SS	8	87.69	93.44	90.47

Sentences with these surface patterns tend to result in numbers of false positive that are significantly larger than with the supervised algorithms. In our experiments, the best-unsupervised results were achieved by using GBT, which returns four false negative and eight false positives.

C. Minimize Human-Labeling Efforts

The SS learning algorithm used in this paper aims to minimize the human-labeling effort while achieving a high prediction accuracy in order to meet the needs of personnel involved in military operations.

When using GBT for the SS framework, the classifier model results in a 90.47% F1 score and needs only eight human labels. Not using the SS framework, a straightforward solution to train a comparably accurate classifier model is to randomly sample some unknown sentences for human labeling.

In Table VI, we compare the accuracy resulting from using different amounts of randomly sampled and hand-labeled sentences. Here, *random 10%* means drawing a random sample of 10% of all sentences from UNCLOS and asking humans to label them. *GBT+SS* is the abbreviation for using GBT with an SS learning framework. We identified comparable recall rates and F1 scores from using the following:

1) a classifier trained on randomly sampled and then hand-labeled 50% (i.e., 1560 sentences) of all sentences in UNCLOS;
2) an SS learner that requires eight nonrandomly selected, hand-labeled sentences (which is about 0.25% of all sentences).

Based on our experience, it takes a person about half a minute to label one sentence. Extrapolating these time costs to train a classifier model based on the random-sampling method that achieves an F1 score of more than 90%, 13 person-hours of labeling efforts would be needed. This is a steep and costly increase over the 4 min of human-labeling effort needed with the SS framework.

VI. Conclusions, Limitations, and Future Work

This paper presents a computational solution for the task of definition acquisition from text collections of military operations law. To approach the definition-detection problem, we converted existing surface patterns into features used for learning. We compared the performance of different supervised learning methods, thus observing a 95.93% F1 score and 96.72% recall rate for the most-accurately performing algorithm. To minimize human efforts in labeling training data, we proposed and implemented an SS solution that balances prediction accuracy and labeling efficiency. The experimental results show a 90.47% F1 score and 93.44% recall rate, which only costs eight human labels.

Several limitations apply: This paper focuses on predicting and extracting explicit definitional sentences. However, there are many implicit definitional sentences that cannot be detected by using the surface patterns we worked with. In our future work, we aim to identify and extract implicit definitions.

Furthermore, as with all machine learning, the resulting model is assumed to deliver similarly accurate predictions when applied to new datasets but might not generalize well across languages, genres, and text data that significantly differ from the training data on other dimensions.

Finally, definition acquisition is just one out of many challenges that people faces when they try to understand legal documents as they relate to military operations. In the future, we will explore logic extraction and representation from respective corpora with the ultimate goal of supporting authorities involved with international conflicts to map their real-world environment to legal constraints and regulations.

Acknowledgment

The authors would like to thank M. W. Bigrigg, G. Davis, and B. Hirshman for their insightful suggestions. They also wish to thank the anonymous reviewers for their comments.

References

[1] S. Blair-Goldensohn, K. R. McKeown, and A. H. Schlaikjer, "A hybrid approach for QA track definitional questions," in *Proc. 10th Text Retrieval Conf.*, 2003, pp. 336–343.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. Workshop Comput. Learning Theory*, 1998, pp. 92–100.

[3] C. Burges, "A tutorial on support vector machine for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[4] Y. Chang, D. Chen, Y. Zhang, and J. Yang, "An Image-based automatic Arabic translation system," *Pattern Recogn.*, vol. 42, no. 9, pp. 2127–2134, 2009.

[5] Y. Chang, H. Xu, and S. Bai, "TREC 2003 question answering track at CAS-ICT," in *Proc. 12th Text Retrieval Conf. Notebook*, 2003, pp. 460–467.

[6] Y. Chang, H. Xu, and S. Bai, "A re-examination of IR techniques in QA system," presented at the 1st Int. Joint Conf. Nat. Lang. Process., Hainan Island, China, 2004.

[7] Y. Chen, M. Zhou, and S. Wang, "Reranking answers for definitional QA using language modeling," in *Proc. 44th Annu. Meeting Assoc. Comput. Ling.*, 2006, pp. 1081–1088.

[8] H. Cui, M. Y. Kan, and T. S. Chua, "Unsupervised learning of soft patterns for definitional question answering," in *Proc. 13th World Wide Web Conf.*, 2004, pp. 90–99.

[9] H. Cui, M. Y. Kan, and T. S. Chua, "Generic soft pattern models for definitional question answering," presented at 28th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, Salvador, Brazil, 2005.

[10] J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002.

[11] A. B. Goldberg and X. Zhu, "Keepin' it real: Semi-supervised learning with realistic tuning," presented at North Am. Chapter Assoc. Comput. Ling. Workshop Semi-Supervised Learn. NLP, Boulder, CO, 2009.

[12] W. Hildebrandt, B. Katz, and J. Lin, "Answering definition questions using multiple knowledge sources," in *Proc. North Am. Chapter Assoc. Comput. Ling. Hum. Lang. Technol.*, 2004, pp. 49–56.

[13] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," presented at Eur. Conf. Mach. Learn., Chemnitz, Germany, 1998.

[14] G. Lampouras and G. Androutsopoulos, "Finding short definitions of terms on web pages," presented at Conf. Empirical Methods Nat. Lang. Process., Singapore, 2009.

[15] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *Proc. AAAI/ICML Workshop Learn. Text Categorization*, 1998, pp. 41–48.

[16] A. McCallum and K. Nigam, "Employing EM and pool-based active learning for text classification," in *Proc. 15th Int. Conf.*, 1998, pp. 359–367.

[17] I. Muslea, S. Minton, and C. Knoblock, "Active + semi-supervised learning = robust multi-view learning," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 435–442.

[18] D. Ravichandran and E. H. Hovy, "Learning surface text patterns for a question answering system," presented at 40th Assoc. Comput. Ling. Conf., Philadelphia, PA, 2002.

[19] A. Steinfeld, P. Quinones, J. Zimmerman, R. Bennet, and D. P. Siewiorek, "Survey measures for evaluation of cognitive assistants," in *Proc. National Inst. Stand. Technol. Perform. Metrics Intell. Syst. Workshop*, 2007, pp. 189–193.

[20] E. M. Voorhees, "Question answering in TREC," in *Proc. ACM Conf. Inf. Knowledge Manage.*, 2001, pp. 535–537.

[21] J. Xu, Y. Cao, H. Li, and M. Zhao, "Ranking definitions with supervised learning method," in *Proc. 14th World Wide Web Conf.*, 2005, pp. 811–819.

[22] J. Xu, A. Licuanan, and R. Weischedel, "TREC2003 QA at BBN: Answering definitional questions," presented at 12th Text Retrieval Conf., Cambridge, MA, 2003.

[23] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions," presented at Workshop Continuum Labeled Unlabeled Data Mach. Learn. Data Mining, Washington, DC, 2003.

[24] Air Force Judge Advocate General's School, *The Military Commander and the Law*. Maxwell AFB, AL: AFJAGS, 2009.

[25] Dept. Defense, *Department of Defense Dictionary of Military and Associated Terms*, Joint Publ. 1-02, 2001.

[26] *The United Nations Convention on the Law of the Sea (A historical perspective)*, United Nations Division for Ocean Affairs and the Law of the Sea, New York, 2009.

[27] *Operational Law Handbook*, The Judge Advocate General's Legal Center and School, Charlottesville, VA, 2009.

Yi Chang (M'07) is currently working toward the Ph.D. degree with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

He is the author or coauthor of more than 25 referred journal and conference publications. His current research interests include machine-learning, natural-language processing, information retrieval, web search, and data mining.

Dr. Chang became a member of the Association for Computing Machinery in 2007.

Jana Diesner (M’11) is currently working toward the Ph.D. degree with the School of Computer Science, Center for Computational Analysis of Social and Organizational Systems, Carnegie Mellon University, Pittsburgh, PA.

Her mission is to span the boundary between natural-language processing and network analysis in order to better understand the coevolution and interplay of the semantics and mechanics of real-world networks. She combines theories and methods from linguistics, social science, machine learning, and computing for this purpose.

Kathleen M. Carley (M’07) received the Ph.D. degree in sociology from Harvard University, Cambridge, MA.

She is currently a Professor of computation, organizations, and society with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, and the Director of the Center for Computational Analysis of Social and Organizational Systems (CASOS). She and the members of CASOS have developed infrastructure tools to analyze large-scale dynamic networks and various multiagent simulation systems. She has authored or coauthored multiple books and more than 100 articles in this area. Her current research interests include cognitive science, social networks, and computer science to address complex social and organizational problems, as well as dynamic-network analysis, computational social and organization theory, adaptation and evolution, text mining, and the impact of telecommunication technologies and policy on communication, information diffusion, disease contagion, and response within and among groups, particularly in disaster or crisis situations.