Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, January 2023

Enhancing Named Entity Recognition and Classification: A Decision Tree Approach

Talasila Ram Kumar
Research Scholar
Shridhar University
Pilani-ChirawaRoad
Pilani, Rajasthan, India

Dr. Sunil Gupta
Professor, DEPT of CSE
Shridhar University
Pilani-Chirawa Road
Pilani, Rajasthan, India

Dr. C. Rama Seshagiri Rao
Pfofessor,Dept OF CSE
Vignana Bharathi Engg College
Hyderabad, India

Abstract

Named-Entity Recognition and Classification (NERC) systems are responsible for assigning semantic labels to phrases that represent named entities, such as individuals, places, and organizations. Typically, these systems rely on two linguistic resources: a recognition grammar and a lexicon containing known names categorized by their respective named-entity types. NERC systems have demonstrated effectiveness in highly specialized domains. However, creating the grammar and lexicon for a new domain is a challenging and time-consuming endeavor. We suggest employing decision tree induction as a potential solution for tailoring a NERC system to suit a specific domain. We propose employing decision trees as NERC "grammatical structures" and generating these trees using machine learning techniques. To validate our approach, we conducted experiments using the C4.5 algorithm to identify names of individuals and organizations associated with management succession events, utilizing data from the sixth Message Understanding Conference. The evaluation results are highly promising, demonstrating that the generated decision tree can outperform a manually constructed grammar.

1 Introduction

Recently, Machine Learning techniques have emerged as a promising solution to a significant challenge in language engineering: the creation of lexical resources. Many practical language engineering systems rely heavily on various lexical resources, including grammars and lexicons. However, using general-purpose resources often proves ineffective because most applications require specialized vocabularies not supported by generic lexicons and grammars. Consequently, substantial efforts are currently dedicated to developing versatile tools that can swiftly adapt to

specific thematic subdomains. This adaptation primarily entails acquiring domain-specific semantic lexical resources.

Named-entity recognition and classification (NERC) involves identifying proper names in text and categorizing them as different types of named entities, such as individuals, organizations, locations, and more. NERC systems assign labels to text phrases based on their corresponding named-entity type, a crucial subtask in information retrieval and data extraction. The categories used for classifying named entities contain semantic information that can vary significantly across different thematic domains. For instance, identifying organization names is relevant in financial news but may not be pertinent in scientific literature. Typical lexical resources in a NERC system include a lexicon in the form of gazetteer lists and a grammar responsible for recognizing entities that are either absent from the lexicon or appear in multiple gazetteer lists. Manual adaptation of these resources to a specific domain is not only time-consuming but sometimes impossible due to the absence of domain experts. Therefore, automating the acquisition of these resources from a training corpus (i.e., textual data) is highly desirable. This article addresses one aspect of this problem, focusing on the acquisition of the NERC grammar.

Research into automatically deriving a NERC grammar from textual data is still in its early stages, with only a few solutions proposed thus far. Conversely, automatic acquisition of recognition and classification models has been extensively explored in the field of machine learning, yielding various successful methods. Among these, the decision-tree induction algorithm C4.5 [1] stands out as the most widely used. C4.5 offers several merits, including its applicability to a range of learning tasks, computational efficiency, and the human-readable format of the generated models, i.e., decision trees. These characteristics make C4.5 a suitable choice for our experiments in learning domain-specific NERC "grammars." In our experiments, we utilized data from the sixth Message Understanding Conference (MUC-6) [2]. The Message Understanding Conferences serve as a primary platform for evaluating new information extraction systems based on a common task. The work described in this article was conducted within the framework of the research project ECRAN-2 which focused on adapting an information extraction system to new thematic domains and languages.

This paper is structured as follows:Section 2 provides an overview of related work on NERC system adaptation, while our approach to the NERC task is explained in Section 3. Experimental

results are presented in Section 4 and Section 5 uses the conclusions drawn from this work to outline future directions.

2 Related Work

As previously mentioned, the NERC task revolves around utilizing gazetteers and named-entity grammars, which require periodic updates when adapting the NERC system to a new domain. Researchers in the field of language engineering have recently shown interest in employing learning techniques to support this adaptation task. Examples of NERC systems that employ supervised learning techniques, whether statistical or symbolic, include Nymble [3], Alembic [4,5], AutoLearn [6], RoboTag [7], and the NYU system for MUC-7 [8]. The approach described here also falls within this category. On the other hand, systems like the NERC system developed for Italian as part of the ECRAN project [9] and the multi-level bootstrapping approach presented in [10] are instances of systems that leverage unsupervised learning.

Nymble [3] utilizes statistical learning to create a Hidden Markov Model (HMM) for recognizing named entities in text. This HMM assigns labels to words, categorizing them into desired named-entity types (e.g., person, organization, etc.) or labeling them as 'NOT-A-NAME.' The HMM states are organized into regions, with one region for each desired type and one for non-named entities. Each region employs a statistical bigram language model, emitting one word upon entering each state. Additionally, states can generate features related to numeric expressions, capitalization, and membership in lists of important words (e.g., company designators, person titles, etc.). Nymble has achieved evaluation results of approximately 90%, and in the MUC-7 competition, it attained 89% recall and 92% precision. The system's success is attributed to its use of appropriate features in word encoding, such as capitalization, and the probabilistic modeling of the recognition process.

Alembic [4] adopts a transformation-based rule learning approach, inspired by Brill's work on part-of-speech tagging [12], for named-entity recognition. This method aims to automatically discover sets of phrase rules in a maximum error-reduction scheme. The learning process begins with an initial labeling function applied to a training corpus. It iteratively evaluates every possible rule to determine its impact on phrase labeling and selects the rule that leads to the greatest reduction in error. Learning continues until a predefined criterion is met, typically when

performance improvement falls below a threshold. Alembic has achieved results of 88% recall and 83% precision, whereas a manually constructed system on the same data achieved 91% recall and 92% precision.

AutoLearn [6] employs the ID3 algorithm [13] to construct decision trees based on hand-tagged training data for detecting the start and end points of specific named entities. Data is converted into five-word tuples, each marked as potentially containing the start or end of a named entity in the middle word. The AutoLearn system achieved limited success in the MUC-6 evaluation, with 47% recall and 81% precision, primarily due to its limited use of lexical resources like gazetteer lists. Improved methods, using the C4.5 algorithm [1], are discussed in [14] and [7]. These methods perform better due to enhanced use of lexical resources and reach comparable or better results.

The NERC system developed for Italian within the ECRAN project [9] employs unsupervised learning to enhance a manually constructed system's performance. Approximately 20% of named entities remain unclassified and are processed by the unsupervised learning algorithm. This algorithm utilizes an untagged learning corpus, a shallow syntactic parser, a "seed" gazetteer, and a dictionary of synonyms to extract elementary syntactic relations (ESLs) from the corpus. ESLs characterize named entities previously classified by the manual system, facilitating the classification of remaining named entities.

The multi-level bootstrapping approach introduced in [10] is partially supervised, utilizing a small number of tagged examples and a larger volume of untagged data. This approach aims to induce information extraction patterns to identify and classify named entities in text. It begins by generating candidate extraction patterns exhaustively, using the AutoSlog system. Additionally, a small number of seed examples for named entities are provided. The most effective pattern for recognizing seed examples is selected and used to expand the set of classified named entities through several iterations, resulting in a dictionary of named entities and corresponding extraction patterns.

Our approach shares similarities with AutoLearn, RoboTag, and the NYU system for Hindi, as we employ the C4.5 algorithm. However, our main distinction lies in the problem representation. Unlike existing approaches, which focus on identifying components of a phrase belonging to a particular NE type, especially their start and end points, we propose a pre-processing step in

which noun phrases are identified separately by a parser. Assuming that NEs are noun phrases, the decision tree can then concentrate on these phrases and classify them into the desired NE types, using a special class for non-NE phrases. This approach mitigates the need for post-processing and enhances interpretability.

3 The NERC Task in MUC-6

To evaluate the C4.5 system, we utilized a portion of the dataset employed in the assessment of systems during the MUC-6 conference [2]. The MUC-6 thematic domain focused on management succession events, encompassing various named entity types, including persons, organizations, locations, dates, times, and monetary values, among others. Generally, it is widely acknowledged (e.g., [17,7]) that identifying and classifying person and organization entities is more challenging. Therefore, our study centers on these two specific entity types. Our dataset consists of 461 organization instances and 373 person instances.

The primary objective of our research is to reduce the human effort required for adapting the Named Entity Recognition and Classification (NERC) system to a specific domain, in this case, management succession events. We employed the VIE NERC system, developed at Sheffield University [18], which utilizes gazetteer lists containing person names, organization names, company designators (e.g., Ltd. and Co.), person titles (e.g., Mr. and MD), and more, in addition to a grammar. The grammar considers tags assigned by referencing the gazetteer lists, as well as the part-of-speech and syntactic properties of words in a phrase. A simple bottom-up chart parser utilizes this grammar to identify significant phrases within the text. Adapting such a system to a particular domain typically involves updating the gazetteer lists and the NERC grammar. In our study, we simplified the adaptation process by focusing solely on the NERC grammar. The construction of gazetteer lists must be completed beforehand. In our investigation, we utilized the following lists: organization (2,559), org_base (55), org_key (80), cdg (94), person (476), title (163), location (2,114), money (101), and time (360), with the numbers in parentheses representing the number of entries in each list.

For C4.5, the data must be transformed into a feature-vector format. In our case, an example corresponds to a Named Entity (NE) phrase, consisting of one or more words, along with contextual information, such as words in close proximity to the NE phrase. Each organization and person instance in the MUC-6 dataset is represented by a feature vector. Two features are

assigned to each word: its gazetteer tag (if available) and its part of speech. Therefore, each vector comprises 28 features, including 14 part-of-speech and 14 gazetteer tags. When the NE phrase contains fewer than 10 words, the remaining features are marked with a special value (a question mark), indicating missing information for the algorithm. Words lacking a gazetteer tag are assigned a special tag called "NOTAG." For instance, if we consider the phrase "… of the Securities and Exchange Commission in the …" with the organization phrase in italics, the corresponding vector is as follows: [IN, NOTAG, DT, NOTAG, NNP, org_key+organisation, CC, organisation, NNP, organisation, NNP, org_base+organisation, ?, IN, NOTAG, DT, NOTAG], where the part-of-speech tags are interpreted as follows: IN: preposition, DT: determiner, NNP: noun phrase, CC: conjunct. The gazetteer tags in this example include organization, org_key, org_base, and NOTAG. Multiple tags for a word are denoted by a plus sign, indicating that the word appears in more than one gazetteer list, as seen with the word "Securities," which is both an org_key and part of an organization ("Securities and Exchange Commission").

In addition to the training examples for person and organization NE phrases, we also constructed a set of negative examples, representing non-NE phrases, to address the dual nature of the NERC task – identifying and classifying NE phrases. By providing the learning algorithm with both person and organization phrases, the resulting decision tree distinguishes between person and organization names. In other words, all phrases up to 10 words are classified as either an organization or a person name. The inclusion of negative examples allows the NERC system to capture patterns that differentiate NE from non-NE phrases. In our study, the negative examples were derived from all noun phrases that are not NE phrases, resulting in a total of 4,333 examples. It's important to note that some non-NE noun phrases may overlap or even encompass each other. Furthermore, they may overlap with or be part of NE phrases. For instance, the noun phrase "George Black's garden" is not a named entity but includes the person name "George Black." Similarly, the phrase "Greek Society for the Protection" is not a named entity but forms part of the organization name "Greek Society for the Protection of Forests." This presents a challenge for the learning algorithm, as it needs to identify what is missing from a phrase (e.g., the word "Forests") rather than what should be included for it to qualify as a named entity.

4 Results

The objective of the presented experiments is to assess the performance of decision trees generated by C4.5 when applied to the Named Entity Recognition and Classification (NERC) task. Additionally, these experiments seek to provide insights into the "grammar" of NERC generated by C4.5, specifically focusing on the information encoded within the constructed decision trees.

To achieve this, we represent the NERC task as a three-class problem, with the three classes being "person," "organisation," and "non-NE" (non-named entity). Consequently, the decision tree produced by C4.5 simultaneously handles both aspects of the NERC task: the identification and classification of named entity phrases. In these experiments, we leverage a unique feature of C4.5, well-suited for handling multi-valued features. This feature allows for the automatic construction of subsets of feature values, as opposed to individually evaluating each feature value whenever it's used in the decision tree. The division of feature values into subsets is guided by the mutual information heuristic employed by C4.5 during decision tree construction. The impact of this feature is a substantial reduction in the branching complexity of the induced tree, making it more interpretable for human understanding.

Within the experiments, various levels of tree pruning are explored, resulting in decision trees of varying sizes. At each size, a 10-fold cross-validation procedure is executed to provide an unbiased estimate of the system's performance on unseen data. Under this evaluation approach, the dataset is divided into ten equally-sized subsets, and the final assessment is an average across ten iterations. During each iteration, nine of the ten data subsets are used to construct the decision tree, while the remaining subset is reserved for evaluation.

The evaluation employs commonly used measures in the field of language engineering, namely recall and precision. Recall quantifies the number of correctly identified items of a specific type (e.g., organizations), divided by the total number of items of that type within the training data. Precision, on the other hand, represents the ratio of correctly identified items of a particular type to all items assigned that same type (e.g., organizations) by the system. Four specific measures are employed in the experiments: recall for organizations, recall for persons, precision for organizations, and precision for persons.

Lastly, for comparative purposes, we can reference the performance of the manually crafted set of rules within the VIE NERC system [18]. The results of this system's performance on our dataset are provided in Table 1.

Recall (o)	*Precision (o)*	*Recall (p)*	*Precision (p)*
71.25%	84.56%	86.78%	93.5%

Table 1. *Performance of the manually constructed set of rules on the whole dataset*

These findings indicate that the outcomes achieved in this study are notably lower when compared to the comprehensive results reported for VIE in MUC-6. The main reason behind this discrepancy lies in the challenges associated with identifying individual and organizational names. VIE's performance also falls notably short of the top-performing system in MUC-6, which managed to attain recall and precision rates exceeding 90%, even for individuals and organizations. It's worth noting that the results displayed in Table 1 demonstrate superior performance in recognizing individual names as opposed to organizational names. This pattern is consistent with the broader trends observed in MUC-6 results, primarily due to the fact that individual names tend to be shorter and are often present in gazetteers or accompanied by personal titles. This makes their identification comparatively easier than identifying organizational phrases, which can be more extensive and encompass various parts of speech and gazetteer entries.

The performance of the decision trees employed in the experiment is illustrated in Figures 1 and 2, which present the average recall and precision results obtained through cross-validation runs for different decision tree sizes, representing various levels of pruning. Figure 1 illustrates the outcomes for organizational phrases, while Figure 2 displays those for individual entities. Each data point on the graph represents the average of 10 values acquired during the corresponding 10-fold cross-validation experiment. Similar to the manually constructed VIE NERC system, the performance for organizational entities falls significantly below that of individual entities. For organizational entities, both recall and precision hover around the 80% mark across different tree sizes. The curves do not exhibit substantial fluctuations as the decision tree size increases, implying that smaller, simpler trees perform just as effectively as larger ones in this task. A comparable trend is observed for individual entities, where the curves for recall and precision

nearly overlap at approximately 90% for decision tree sizes exceeding 50 nodes. Smaller trees seem to miss some individual entities, resulting in lower recall values.

In summary, the performance observed is approximately 80% in terms of recall and precision for organizational entities and around 90% for individual entities. When compared to the manually constructed VIE system, precision is lower by approximately 3.5 percentage points for organizational entities and 2.5 percentage points for individual entities. However, recall is higher by roughly 10 percentage points for organizational entities and remains consistent for individual entities. Overall, the automatically generated trees yield results that are favorable when compared to those of the manually constructed grammar.

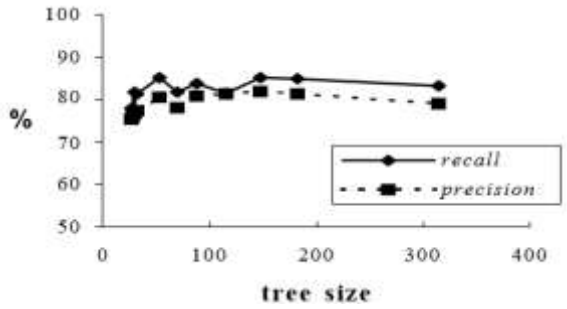
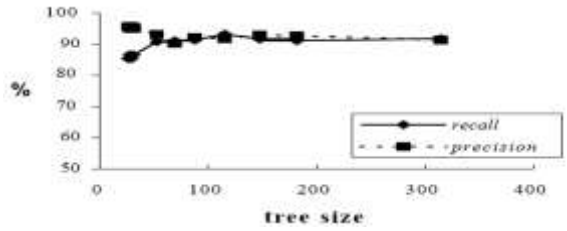


Figure 1. *Recall and precision results for organizations*

Moreover, our findings demonstrate a level of similarity to the outcomes reported in the RoboTag study [7]. Regrettably, a direct comparison with the NYU system isn't feasible since we haven't assessed our system's performance on Hindi texts.



The noteworthy effectiveness of our approach becomes apparent, considering that it eliminates the need for a post-processing stage. Specifically, the decision trees yield NE phrase classifications directly, making them readily interpretable by humans. In fact, they can be conveniently translated into IF-THEN-ELSE rules.

Recall (o)	*Precision (o)*	*Recall (p)*	*Precision (p)*	*Tree Size*
89.6%	86.6%	93.0%	95.6%	58 nodes

Table 2. *Performance of the representative classifier on the whole dataset.*

Figure 3 illustrates a set of rules derived from a decision tree comprising 58 nodes, generated during our experimentation. At this scale, the tree's performance already attains the levels mentioned previously. In particular, the tree's performance, as depicted in Figure 3, across the entire dataset (including both training and test data) is detailed in Table 2.

The classifier depicted in Figure 3 reveals intriguing patterns, one of which involves recognizing that a person's name is typically preceded by a title (as indicated by Gtag(-1) being within the set {title, org_key+title}). Of particular note are the rules that identify sub phrases within organization and personal names as non-named entities (NE). These rules account for special cases involving incomplete named entities.

An example of such a rule pertains to situations where the first word in a phrase could potentially be associated with an organization (Gtag(1) IN {location+organization, org_key+organization, organization+person}), but the following word in the phrase is identified as a company designation (cdg), an organization, or a person. In such cases, the phrase is classified as non-NE, as it forms only part of an organizational entity.

```
Representative classifier: (abbreviations: POS=part-of-speech tag, Gtag=gazetteer tag)

IF Gtag(-1) IN {title, org_key+title} THEN person

ELSEIF Gtag(-1) IN {NOTAG, currency_unit, date, location, org_key+organization,
        organization, person} THEN:

    IF Gtag(1) = NOTAG THEN:

        IF POS(1) IN {NNP, VBG} THEN:

            IF POS(+2) IN {RP, VB, WP} THEN person

            ELSEIF POS(+2) IN {CD, MD, NN, NNS, RB, TO}

                AND POS(-1) IN  {CC, NN, PERIOD, SYM, VB, VBD, VBZ}

                THEN person

            ELSE organization

        ELSE non-NE

    ELSEIF Gtag(1) IN {cdg, govern_key, location, location+title, org_base, org_key, title} THEN:IF
        POS(1) IN {NNP, VBG} AND POS(-1) IN {DT, JJ} THEN organization

        ELSE non-NE

    ELSEIF Gtag(1) IN {location+organization, org_key+organization, organization+person}
        THEN:IF Gtag(+1) IN {cdg, organization, person} THEN non-NE

        ELSEIF Gtag(+1) = NOTAG THEN:

            IF POS(+1) IN {CC, COMMA, DT, IN, JJ, JJR, NN, NNS, POS, SYM, TO, VB, VBD,
            VBZ, WP}

                THEN:

         IF POS(4) IN {CC, IN, JJ, NN, NNS, VBD, VBZ}

         AND POS(2) IN {COMMA, IN, NN, NNS, POS}
```

Figure 3. A representative classifier, corresponding to a decision tree containing 58 nodes.

5 Conclusion

In this article, we assessed the performance of the C4.5 algorithm when applied to the task of training decision trees for the identification and categorization of named entities in textual data. This approach significantly reduces the labor required to tailor a Named Entity Recognition and Classification (NERC) system to a particular domain. The experiments conducted yielded several valuable insights regarding the applicability of C4.5 in this context:

An important finding is that the performance of named entity recognizers generated by C4.5 stands up favorably when compared to manually constructed recognizers employing the same

lexicon. Hence, employing C4.5 for adapting a NERC system is highly recommended. Moreover, the recognizers created by C4.5 tend to be straightforward and can be easily translated into a concise set of understandable rules. The simplicity of these rules is further enhanced through a specialized feature in C4.5, which permits the automatic sub-setting of feature values for multi-valued attributes. The classification rules generated by C4.5 can be scrutinized and refined by human experts.

The results we obtained are in line with similar systems utilizing the C4.5 algorithm for NERC. Additionally, our representation of the NERC problem obviates the need for a post-processing stage, which is a common component in all decision tree-based systems. Consequently, the decision tree directly furnishes the NERC rules.

An intriguing avenue for future research involves a comparative assessment of alternative learning methods. The initial set of contenders may include learning methods employing the same feature-vector representation as C4.5, such as AQ15 and CN2. Alternatively, methods that explicitly engage in grammar induction, as well as unsupervised learning methods, warrant further exploration. The latter could potentially reduce human involvement in the learning process, particularly by eliminating the necessity for manual data tagging while maintaining recognition performance.

Reducing human effort could also be achieved by automating the construction of the lexicon used in NERC, a direction we are currently pursuing. Ultimately, our objective is to develop a methodology for training NERC systems that match the performance of the best manually crafted systems. To this end, we are actively working on enriching our training data representation by incorporating more informative features related to the constituent words of named entities.

In summary, the findings presented in this paper demonstrate the effectiveness of utilizing learning methods to efficiently customize NERC systems for specific domains. Consequently, we view this work as a significant step toward creating NERC systems that can readily adapt to a wide range of real-world applications.

References

1. Jain A., Arora A.: Named entity system for tweets in Hindi language, International Journal of Intelligent Information Technologies (IJIIT), vol. 14(4), pp. 55–76, 2018.

2. Hennig L., Truong P.T., Gabryszak A.: MobIE: A German Dataset for Named Entity Recognition, Entity Linking and Relation Extraction in the Mobility Domain, arXiv preprint arXiv:210806955, 2021.
3. Ekbal A., Saha S., Sikdar U.K.: On active annotation for named entity recognition, International Journal of Machine Learning and Cybernetics, vol. 7(4), pp. 623–640, 2016.
4. Fu R., Qin B., Liu T.: Generating Chinese named entity data from parallel corpora, Frontiers of Computer Science, vol. 8(4), pp. 629–641, 2014.
5. Day, D., Robinson, P., Vilain, M., and Yeh, "A. Description of the ALEMBIC system as used for MUC-7." In [11], 1998. Cowie, J. "Description of the CRL/NMSU System Used for MUC-6." In [2], 1995.
6. Bennett, S.W., Aone, C. and Lovell, C. "Learning to Tag Multilingual Texts Through Observation." In *Proceedings of the Second Conference on Empirical Methods in NLP*, pp. 109-116, 1997.
7. Sekine, S., "NYU: Description of the Japanese NE System used for MET-2." In [11], 1998.
8. Cuchiarelli, A., Luzi, D., and Velardi, P. "Automatic Semantic Tagging of Unknown Proper Names." In *Proceedings of COLING-98*, Montreal, 1998.
9. Riloff, E. and Jones, R., "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping." In *Proceedings of the National Conference on Artificial Intelligence*, pp. 474-479, 1999.
10. Defense Advanced Research Projects Agency. Proceedings of the Seventh Message Understanding Conference (MUC-7), Morgan Kaufmann, 1998.
11. Brill, E. "A corpus-based approach to language learning." *PhD Dissertation*, Univ. of Pennsylvania, 1993.
12. Quinlan, J.R. "Machine Learning: Easily Understood Decision Rules." In *Computer Systems that Learn*, eds. Weiss, S.M. and Kulikowski, C.A., Morgan Kaufmann, 1991.
13. Gallippi, A., "Recognizing Names Across Languages." In *Proceedings of the 16th International Conference on Computational Lingustics (COLING),* 1996.
14. Sekine, S., Grishman, R. and Shinnou, H., "A Decision Tree Method for Finding and Classifying Names in Japanese Texts." In *Proceedings of the Sixth Workshop on Very Large Corpora,* 1998.

15. Riloff, E., "Automatically Constructing a Dictionary for Information Extraction Tasks." In *Proceedings of the National Conference on Artificial Intelligence*, pp. 811-816, 1993.

16. Appelt, D.E., Hobbs, J.R., Bear, J., Israel, D., Kameyama, M., Kehler, A., Martin, D., Myers, K., Tyson, M. "SRI International FASTUS System MUC-6 Test Results and Analysis." In [2], 1995.

17. Humphreys, K., Gaizauskas, R., Cunningham, H., and Azzam, S. VIE Technical Specifications. Department of Computer Science, University of Sheffield, 1997.

18. Michalski, R. S., Mozetic, I., Hong, J. and Lavrac, N., "The multi- purpose incremental learning system AQ15 and its testing application to three medical domains." In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1041-1045, 1986.

19. Clark, P. and Niblett, T., " The CN2 algorithm." *Machine Learning*,
3(4), pp. 261-283, 1989.