In Proceedings of Third International Conference on Knowledge Discovery and Data Mining (KDD-97)

Learning to Extract Text-based Information from the World Wide Web

Stephen Soderland
Dept. Computer Science & Engineering
University of Washington
Seattle, WA 98195-2350
soderlan@cs.washington.edu

Abstract

There is a wealth of information to be mined from narrative text on the World Wide Web. Unfortunately, standard natural language processing (NLP) extraction techniques expect full, grammatical sentences, and perform poorly on the choppy sentence fragments that are often found on web pages.

This paper[1] introduces Webfoot, a preprocessor that parses web pages into logically coherent segments based on page layout cues. Output from Webfoot is then passed on to CRYSTAL, an NLP system that learns text extraction rules from example. Webfoot and CRYSTAL transform the text into a formal representation that is equivalent to relational database entries. This is a necessary first step for knowledge discovery and other automated analysis of free text.

Information Extraction from the Web

The World Wide Web contains a wealth of text information in the form of free text. Until a text extraction system transforms it into an unambiguous format, much of this information remains inaccessible to automated knowledge discovery techniques.

Successful text extraction has been primarily limited to web pages that include tables of information. A system can extract information with high reliability based on the HTML tags used to delimit table entries (Doorenbos *et al.* 1997) (Kushmeric *et al.* 1997). Unfortunately, such systems cannot handle the large proportion of text data that is in narrative form.

Considerable progress has been made in natural language processing text extraction systems (Weischedel 1995) (Grishman 1995) (Krupka 1995). However, NLP techniques typically expect the text to be in the form of full, grammatical sentences. What is found on the web is often a series of brief sentence fragments such as the excerpt from a National Weather Service web page in Figure 1.

A new parser for web pages, Webfoot, demonstrates that NLP techniques can be extended to extracting information from non-grammatical text on the web. The information to be extracted is the *relationships* between individual facts. It is not enough to identify isolated facts, which can be done by a simple key word search. The domain used in this paper is weather forecast web pages, extracting weather conditions associated with a day and location. The output is represented as case frames with slots for Day, Conditions, High temperature, Low temperature, and Location.

A typical NLP information extraction system parses each sentence, then applies rules based on the syntactic relation of phrases within a sentence. Such a system will find no useful syntactic clues in the text in Figure 1. Worse yet, a system that treats each phrase ending with a period as a separate sentence will have difficulty associating "CHANCE OF RAIN 80 PERCENT" with "TONIGHT" rather than with "THURSDAY".

Webfoot takes a web page source text as input, applies rules based on page layout cues, and divides the text into logically coherent segments that are passed to an NLP information extraction system. Webfoot handles a wide range of web page styles, including pages whose layout is indicated by HTML tags or by blank lines and white space, and pages with information in tabular or narrative format. This greatly expands the range of text data that can be extracted automatically from web pages.

The NLP system used in these experiments is CRYSTAL, which learns domain-specific text extraction rules from examples (Soderland *et al.* 1995) (Soderland 1997). The remainder of this paper describes the Webfoot and CRYSTAL systems and presents empirical results for the domain of weather forecast web pages. The combination of Webfoot and CRYSTAL achieve surprisingly good performance for an NLP system operating without the aid of syntactic knowledge. This opens the way for automatic analysis of a class of text data that has been largely inaccessible.

Webfoot: Parsing Web Page Layout

Webfoot uses page layout cues to divide a web page source text into sentence-length segments of text as the first step in an information extraction system. Ideally, these text segments should group together logi-

```
<HEAD>
<TITLE>Forecast for NY072</TITLE>
</HEAD>
<BODY>
<pre>
BRONX-KINGS (BROOKLYN)-NASSAU-NEW YORK (MANHATTAN)-QUEENS-
RICHMOND (STATEN IS.)-
300 PM EST WED FEB 26 1997

.TONIGHT...CLOUDY WITH OCCASIONAL LIGHT RAIN. LOW IN THE MID 40S.
WIND SOUTHWEST 10 TO 15 MPH. CHANCE OF RAIN 80 PERCENT.
.THURSDAY...MOSTLY CLOUDY...WINDY AND MILD WITH A 30 PERCENT CHANCE
OF SHOWERS. HIGH 60 TO 65. WIND SOUTHWEST INCREASING TO 20 TO 30 MPH
WITH HIGHER GUSTS DURING THE AFTERNOON.
.THURSDAY NIGHT...PARTLY CLOUDY. LOW 40 TO 45.
.FRIDAY...MOSTLY SUNNY. HIGH 50 TO 55.
```

Figure 1: Source text from a National Weather Service forecast web page

cally related facts and should separate unrelated facts. The notion of a coherent text segment depends on what facts and relationships are of interest to a given domain. For weather forecast web pages, a segment should include all the weather conditions related to a given day as shown in Figure 2 and should not contain multiple days or include conditions for other days.

```
<segment>
Field(1): <HEAD> <TITLE> Forecast for NY072 </TITLE>
Field(2): </HEAD> <BODY>
</segment>

<segment>
Field(1): <pre>
Field(2): BRONX - KINGS ( BROOKLYN )- NASSAU -
          NEW YORK ( MANHATTAN )- QUEENS -
          RICHMOND ( STATEN IS.)-
Field(3): 300 PM EST WED FEB 26 1997
</segment>

<segment>
Field(1): . TONIGHT ...
Field(2): CLOUDY WITH OCCASIONAL LIGHT RAIN .
Field(3): LOW IN THE MID 40S .
Field(4): WIND SOUTHWEST 10 TO 15 MPH .
Field(5): CHANCE OF RAIN 80 PERCENT .
</segment>
...
<segment>
Field(1): . FRIDAY ...
Field(2): MOSTLY SUNNY .
Field(3): HIGH 50 TO 55 .
</segment>
```

Figure 2: The sample text as segmented by Webfoot

Figure 3 summarizes the tags and other text cues that Webfoot uses to delimit segments and to further break segments into fields. This particular set of delimiters should be seen as a snapshot of a system under development, rather than a fixed set of rules.

Webfoot begins by breaking the web page source text into segments on level 1 delimiters. If a segment has fewer than twenty words (not counting HTML tags), the higher level delimiters are used to break the segment into fields. If the segment has twenty or more words, level 2 delimiters are used as segment breaks.

Domain-independent delimiters:

Level 1	start:	<html, <table, <ul, <pre>
	end:	</html>, </table>, , </pre>
Level 2	start:	<tr
	end:	</tr>
Level 3	start:	, , <hr>, S
	end:	, , </td>
Level 4	start:	<option, <hD
	end:	 , </hD>, </title>, </th>, "$.S$"
Level 3p	start:	line beginning with word(s) followed by ":", line with tabs or multiple spaces
Level 3p	end:	blank line, line with fewer than 40 characters
Level 4p	end:	"$.S$", tabs or multiple spaces

Weather forecast domain delimiters:

Level 3p	start:	line with a weekday followed by ":" "-" or "...", sentence beginning with a weekday, line beginning with "."

S stands for whitespace, and D stands for digit

Figure 3: Delimiters used by Webfoot to parse web page layout

If a segment still has more than twenty words, then level 3 delimiters are used as segment breaks, otherwise as field breaks. In any case, level 4 delimiters are used as field breaks. Pre-formatted sections and pages without HTML tags use delimiters 3p and 4p rather than 3 and 4.

In addition to domain-independent rules, Webfoot may be tailored to the writing style of web pages in a particular domain with the addition of domain-specific delimiters. Three additional delimiters were added for the weather forecast domain to force a new segment when a new day of the week was mentioned and for a National Weather Service convention of beginning bulleted items with a period.

CRYSTAL: Learning Extraction Rules

CRYSTAL is an NLP system that automatically induces a set of domain-specific information extraction rules from training examples. The input to CRYSTAL

is a set of *instances* that are produced by Webfoot or some other sentence analyzer. Each instance is a text segment, divided into fields. A syntactic analyzer would label these fields as "subject", "verb", "object", and so forth. Webfoot simply calls them "field".

An additional input to CRYSTAL is a semantic lexicon used to look up the word sense of individual words in the text. This allows CRYSTAL to create rules that apply to broad classes of words, which is critical for leveraging broad coverage from a limited amount of training. For the weather forecast domain, a semantic lexicon was created by hand consisting of 86 words with the semantic class Weather_Condition ("cloudy", "fair", "precipitation", etc.) and 42 words with the semantic class Time or Day.

CRYSTAL rules, called *concept definitions* have a set of constraints that apply to fields in an instance. These may require the field to include particular semantic classes or terms. A term may be a word, punctuation, or HTML tag.

Some of the fields in the concept definition are designated as extracting one or more slots of the target concept. If all the constraints in a concept definition are met, CRYSTAL creates a case frame with fields of the instance filling slots in the case frame as specified in the concept definition.

Figure 4 shows a concept definition that was induced from a set of National Weather Service web pages. This concept definition has constraints on three fields. One field must include the semantic class Weather_Condition and also a period. Another field must include the semantic class Day and both a period and "...". A third field must include the word "high" and a period. If all three constraints are met, then CRYSTAL extracts Conditions, Day, and High from the fields indicated in the concept definitions.

```
Concept-type Forecast              ID: 459
Status: GENERALIZED
Constraints:
    Field::                        (extract Conditions)
        classes:   Weather_Condition
        terms:     "."
    Field::                        (extract Day)
        classes:   Day
        terms:     ".", "..."
    Field::                        (extract High)
        terms:     "HIGH", "."
      Coverage: 94                 Errors: 1
```

Figure 4: A CRYSTAL concept definition for Day, Conditions, and High

CRYSTAL uses a machine learning covering algorithm similar to Michalski's AQ algorithm (Michalski 1983) and Clark and Niblett's CN2 algorithm (Clark and Niblett 1989). It is a supervised learning method that requires manually annotated training – texts in which each reference to target concepts of the domain has been tagged. CRYSTAL begins with the most restrictive concept definitions that cover each positive training instance. Concept definitions are then generalized by unifying similar definitions.

Empirical Results

Webfoot and CRYSTAL were tested on a domain of weather forecast web pages. Three weather sources were tested that represent widely divergent styles of web pages. The CNN Weather Service has automatically generated pages with extensive use of HTML tags. The National Weather Service (NWS) presents information in series of related sentence fragments, but uses different page layout styles for different regional weather centers. The Australian Bureau of Meteorology (Aus) has web pages without HTML tags and with no consistency in page layout for different regions. A corpus of twenty web pages were annotated for each weather service, two pages each from ten cities or regions.

The metrics used are *recall* and *precision*. Recall is the percentage of positive instances that were identified by the system. Precision is the percentage correct of the instances reported as positive. Rules for each combination of case frame slots were tested separately. If the rule base extracts this combination of slots from a test instance, and each extracted field has the appropriate annotation, this counts as one correct. If any of the extracted fields lack the proper annotation, it counts as an error.

Table 1 shows results for the weather forecast domain. These are averages of fifty random partitions into 90% training and 10% blind test set. The columns labeled NWS-1 and Aus-1 used Webfoot with only domain-independent rules. For NWS-2 and Aus-2, three domain-specific delimiters (tailored to NWS) were added to Webfoot. CNN used only domain-independent rules.

The CNN web pages have such a high regularity that 100% reliable rules can be learned from as few as two training documents. The National Weather Service pages are less rigidly formatted and present information in sentence fragments, but CRYSTAL was able to learn reliable rules that extract over 90% of the information with precision over 90% for combinations of slots that include weather conditions.

Web pages from Australia were so varied that training from one weather station was often little help in learning rules for another weather station. For higher performance, the training data should include multiple pages for each weather station in the test set.

Extracting information about Location showed the lowest performance. CRYSTAL was provided with a semantic lexicon for words with semantic class Weather_Condition, Time, or Day, but there was no corresponding list of city names and geographical terms. CRYSTAL can compensated for this somewhat by learning location names one at a time as term constraints, if the training mentions a location multiple times.

Table 1: Performance in the Weather Forecast Domain

	CNN		NWS-1		NWS-2		Aus-1		Aus-2	
Concept	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre
Day,Conditions,High,Low	100.0	100.0	28.0	5.0	90.3	95.9	52.0	94.0	34.1	85.5
Day,Conditions,High	100.0	100.0	18.6	25.4	92.0	95.4	26.9	83.3	63.0	88.3
Day,Conditions,Low	100.0	100.0	9.3	15.1	90.7	92.6	67.3	95.3	47.1	94.2
Day,Conditions	100.0	100.0	21.1	64.0	96.5	92.1	34.6	83.3	46.3	83.2
Day,Location	100.0	100.0	50.2	44.8	59.1	97.0	23.1	56.4	18.3	64.9
Location	100.0	100.0	55.5	49.4	63.1	97.0	23.8	81.4	17.2	75.0
% Correct Segments	100.0		20.9		98.3		81.8		95.8	
% Lump errors	0.0		79.1		1.7		14.9		0.9	
% Split errors	0.0		0.0		0.0		3.3		3.3	

Table 1 also shows the percentage of correct segmentation[2] for each of the weather sites and each version of Webfoot. "Lump errors" are when unrelated information or multiple days' weather is included in the same segment. "Split errors" are when related information is split between two or more segments. Without the domain-specific rules, Webfoot often ran together entries for multiple days, especially in NWS web pages. This had a serious impact on recall and precision.

Conclusions

Webfoot and CRYSTAL allow automatic information extraction from a class of web page text data that has been largely inaccessible to automated systems. Text in non-grammatical sentence fragments as well as text in tabular format are parsed by Webfoot into coherent text segments based on page layout cues. CRYSTAL then learns domain-specific rules for information extraction. High performance can be obtained even though CRYSTAL was originally designed to rely on syntactic information within full sentences.

Extraction rules for highly structured tables can be learned from as few as two training documents. The performance of Webfoot on this domain suggests that a set of domain-independent rules are sufficient to parse web pages that make extensive use of HTML tags.

Free text narrative and web pages without HTML tags pose a harder problem, and require several annotated training examples for each web site. Webfoot may need a small number of domain-specific rules, such as those testing for days of the week to begin a new segment in weather forecast pages. In addition, CRYSTAL needs a semantic lexicon that lists words of semantic classes relevant to the domain.

Webfoot and CRYSTAL create a formal representation of the text that is equivalent to relational database entries. This provides unambiguous input to later processing, such as classifying individual texts, summarizing data from a large collections of texts, and discovering trends and relationships that span texts. The current experiments are limited to developing and testing the text extraction tools. The next step is to incorporate Webfoot and CRYSTAL as components in a full knowledge discovery application.

[2] of segments containing relevant information

Acknowledgments: This research was funded in part by Office of Naval Research grant 92-J-1946, by ARPA / Rome Labs grant F30602-95-1-0024, by a gift from Rockwell International Palo Alto Research, and by National Science Foundation grant IRI-9357772. CRYSTAL was provided by the NLP Laboratory, University of Massachusetts Computer Science Department, Amherst, Massachusetts.

References

Doorenbos, R., Etzioni, O., and Weld, D. A Scalable Comparison-Shopping Agent for the World-Wide Web In *Proceedings of the First International Conference on Autonomous Agents*, 39-48, 1997.

Clark, P. and Niblett, T. The CN2 Induction Algorithm. *Machine Learning*, 3, 261-283, 1989.

Grishman, R. The NYU System for MUC-6 or Where's the Syntax? In *Proceedings of the Sixth Message Understanding Conference*, Morgan Kaufmann Publishers, 167-175, 1995.

Kushmerick, N., Weld, D., Doorenbos, R. Wrapper Induction for Information Extraction. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997.

Michalski, R. S. A Theory and Methodology of Inductive Learning. *Artificial Intelligence*, 20, 111-161, 1983.

Soderland, S., Fisher, D., Aseltine, J., Lehnert, W. CRYSTAL: Inducing a Conceptual Dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1314-1321, 1995.

Soderland, S. Learning Text Analysis Rules for Domain-specific Natural Language Processing. Ph.D. thesis, technical report UM-CS-1996-087 University of Massachusetts, Amherst, 1997.

Krupka, G. SRA: Description of the SRA System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*, Morgan Kaufmann Publishers, 221-236, 1995.

Weischedel, R. BBN: Description of the PLUM System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*, Morgan Kaufmann Publishers, 55-70, 1995.