# Rethinking Text Segmentation Models: An Information Extraction Case Study

**Christopher D. Manning**

## Abstract

I examine the use of text segmentation methods on semi-structured text within an information extraction application, arguing that: hierarchical models of discourse structure are necessary, that discourse relationships are not, that repetition can indicate cohesion *or* a topic change, and that generalized cue phrases are underutilized. I present results and discussion of the evaluation of two models.

This paper discusses one aspect of an information extraction system for real estate classified ads: how to handle advertisements that advertise multiple properties. Examples of such advertisements appear in Figures 1 and 2. The system described is in production use.[1] The contributions of this paper include:

- Drawing connections and contrasts between approaches to segmentation within the information extraction literature and the related literature on *text segmentation*
- Questioning the utility – for even practical 'language engineering' purposes – of linear segmentations of texts
- Showing that the notion of *topic*, used as a criterion for segmentation, has not been well defined
- Arguing that, for important practical notions of topic, the lexical cohesion metrics most commonly used for text segmentation do not give the desired results and one often wants repulsion
- Reasserting the importance of using a generalized notion of cue phrases
- Developing an information extraction system which demonstrates a practical application of text segmentation
- Providing a task-based evaluation of a text segmentation system, within the context of a discussion of appropriate evaluation methods for text segmentations.

After some introductory material, this paper first discusses theoretical and practical issues in discourse structuring. It then briefly discusses the problem domain and the information extraction system used, and finally discusses issues of performance evaluation, including providing results for the systems discussed in the paper.

Within information extraction (IE), a distinction is often made between single-slot vs. multiple-slot extraction (Soderlandin press). Single-slot systems (e.g., RAPIER, (Califf and Mooney1997); AutoSlog, (Riloff1993) just extract isolated facts from a text. Within some domains, single-slot extraction is adequate; essentially, there must be only one event of interest per document. Multiple-slot systems (e.g., CRYSTAL and WHISK (Soderland et al.1995, Soderlandin press)) group related information into a multi-slot feature-value frame, and can extract several frames from a document. Soderland(in press) also makes distinctions between structured text (things like tabular information on web pages), semi-structured text (things like conference announcements or advertisements, which are basically text in a human language, but which tend to have a certain format and a somewhat telegraphic style), and free text (newspaper articles, expository text, and so on). The extracting and then merging approach is perhaps most suited to 'free text' domains, as have been explored in the MUC conferences. For semi-structured text, but perhaps also for free text, an appealing alternative is to apply work in discourse

[1] It runs in the background of http://www.newsrealestate.com.au/.

structure and text segmentation. The working hypothesis is that a coherent unit of discourse will describe a single event. The system described here adopts this design. Following the phases of tokenization and semantic typing, ads are segmented. Information extraction is then applied to a discourse structure tree. The next part of this paper examines what amount of discourse structure is needed to support such an approach.

## Discourse structure

### Hierarchical structure

It is standardly assumed within theoretical work on discourse structure that discourse has a hierarchical tree structure (Grosz and Sidner1986, Mann and Thompson1987), and an attempt to induce such hierarchical structure is presented in (Morris and Hirst1991). However, within empirically-based work which attempts to automatically label discourse structure, this assumption has been largely abandoned, and discourse structuring is seen as purely a task of linear segmentation – of cutting lengths off the string of discourse. Hearst(1994) notes that the hierarchical view of discourse is standard but proposes that a linear segmentation is sufficient for the domains of interest to her (segments for use in word sense disambiguation and for limiting search and returning context in information retrieval systems). Passonneau and Litman(1993) move to a linear model more for practical reasons. They note that:

> To simplify data collection, we did not ask subjects to identify the type of hierarchical relations among segments illustrated in Figure 1 [because] ... subjects found it difficult and time-consuming to identify non-sequential relations. ... this is consistent with previous findings (Rotondo1984) that non-linear segmentation is impractical for naive subjects in discourses longer than 200 words. (Passonneau and Litman1993:149)

It is, however, worrisome that the motivating example in their paper for the utility of discourse structuring – coreference resolution within the dialogue referred to in the quote (not reproduced here) – depends completely on the recognition of *hierarchical* discourse structure.

Let us now look at the segmentation of real estate ads, such as those in Figures 1 and 2. Within this domain, there is just no difficulty in determining the correct hierarchical segmentation. It

```
ABSOLUTE
RIVER FRONT
ACREAGE
50 ACRES $119,000
80 ACRES $139,000
Unique secluded Freehold
riverfront acreage - natural
riverland beauty.
River gums, willows, natural
lagoon, birdlife and sandhills.
Huge cliffs opposite. Ideal ski
waters, 90 mins Adelaide.
Opportunity to build your own
riverfront home (subject to
approvals).
MANNUM
Family or retirement home.
Brand new and ready for you.
Spacious 3 bedrooms, family
room. Carport u.m.r. Fully
fenced. $105,000.
WALKER FLAT
Riverside holiday home above
flood level with access to river
and jetty. Comfortable 3
bedrooms, boatshed and
carport. $69,500.
Adrian Davis (08) 8569 2120
WESTON RAINE & HORNE
96 Randell St. Mannum
```

(a)

```
Escape To Paradise
ON PORT STEPHENS
AT TEA GARDENS
ON THE MYALL RIVER
BRAND NEW HOMES
Only two hours North of Sydney
nestling on the Myall River is a
beautiful little town of thoughtful
caring people.
A place unspoiled by heartless
developers - a place where brand
new homes can be purchased only
minutes from shops & beach.
Brand new Home/ Land Packages
complete with quality carpet,
lightfittings, driveway, turf,
fencing, clothesline & letter box.
4 Bed, ensuite, Dble Gge$137,844
4 Bed, ensuite, Dble Gge $148,000
4 Bed, ensuite, rumpus $162,950
In the meantime let us send you
some plans and other stuff Phone
Karen on toll Free ``Escape Line''
MYALL QUAYS
1800 242 321
```

(b)

Figure 2. House for sale advertisements

```
ACRES 3br $180, $230. 4br $210 BONDI J 2br $300, 3br furn, $420
PADDO 4br $390, 2br furn, $360 GLEBE 5br $440, 2br $270, $250
NEWTOWN 3br $275 or 4br $310 P'MATTA 1br 1/2hse $130 3br $185
SURRY HILLS 4br $450, 2br $200 RYDE 3/4br $245, 2br yd $190
COOGEE 3/4br $390 or 4br $420 BALMAIN 3br $365, 1br yd $220
NEUT BAY 2br yd $265, 4br $450 COOGEE 1 br $190 or 2/3br $295
BONDI 1/2 br $195, 3br yd $450 H.R. Licensed FEE 9516-3211
```

Figure 1. A rental advertisement

1. `A bomb exploded last night in downtown Bogota.`
2. `Three people were injured, none seriously.`
3. `This comes after an incident 3 days ago when a police station was damaged.`
4. `Two vehicles were also destroyed.`
5. `Yesterday's attack was condemned by officials.`
6. `Police are searching for two hooded men.`

Figure 3. An extract from a MUC-4 newspaper article, illustrating the need for hierarchical structure

is obvious which parts of an ad are describing different properties, and which parts hierarchically have scope over several properties.[2] So, practical difficulties in getting reliable hierarchical segmentations do not exist. Regarding the argument of what is necessary for language engineering tasks, linear segments are not sufficient for good IE results. If a suburb name, inspection time, or contact number has scope over several properties, then it is vital to know. Otherwise the IE system will perform poorly when evaluated in terms of its success in filling slots in per-property feature-value frames. On the other hand, there seems little practical need to relate the segments of the text via explicit discourse relations such as *elaboration* or *justification*.

One could feel that this need for hierarchical structure exists only in the somewhat stylized and compressed domain of newspaper advertisements. But I believe that the need is widespread throughout places where IE might be applied. As an example from a different domain, consider the text in Figure 3 from the MUC-4 terrorist incident domain (this is the running example in (Crowe1995)). In this dialog, sentences 3 and 4 are a text segment that is embedded in the segment described in setences 1, 2, 5, 6. Good performance for an IE system requires noticing that sentences 5 and 6 contain more information about the event described in sentences 1 and 2 (such as that the perpetrators are believed to be "two hooded men").

[2] I have not done a formal study with human segmenters to verify this claim, but I believe it to be quite uncontroversial, as the reader can confirm from the examples in this paper.

## Signifiers of a segment: what is a "topic"

The two basic techniques that have been used for segmentation are cue phrases (e.g., (Passonneau and Litman1993, Litman and Passonneau1995) following (Grosz and Sidner1986)), and lexical cohesion (Morris and Hirst1991, Kozima1993, Hearst1994, Richmond et al.1997). True cue phrases (such as *for example* or *basically*) are valuable in processing spoken discourse. It seems clear that they are not much use in segmenting advertisements, but nevertheless certain elements such as suburbs and prices can perhaps be regarded as generalized cue phrases, a topic to which I return below.

The central idea of systems based on lexical cohesion (Morris and Hirst1991) is that if the text continues to use similar words, then it is probably still talking about the same topic. Looking for just repeated words is insufficient and so most systems, such as (Morris and Hirst1991), use thesauri or semantic nets so that cohesion is being evaluated at a semantic level, rather than at a lexical level. Is lexical cohesion a good tool for segmentation? To answer this, we need a better idea of what the desired segments are. Much of the existing literature on text segmentation is quite vague in defining the desired segments (Beeferman et al.1997). Hearst(1994) suggests segment boundaries should be where "a relatively large set of active themes changes simultaneously". Kozima(1993) suggests: "a scene in a movie which describes the same objects in the same situation." Beeferman et al.(1997) adopt the empirical definition that a segment boundary is the article boundaries between news reports in newswire corpora. Lexical cohesion works well in this domain, because successive stories are almost always about completely different topics. This contrasts with the actual formatting of a newspaper, where stories on the same topic are normally grouped. On days with big stories (Iraq bombing, Clinton impeachment, ...) the entire front page may be devoted to one topic, but will contain multiple articles. In such circumstances, there is likely to be great lexical cohesion between everything on the front page, but, nevertheless, the articles represent different movie scenes. Similarly, within real estate ads, every ad talks about bedrooms, garages, locations and prices. There is great lexical cohesion throughout the entire real-estate section. But, nevertheless, we wish to segment it into vignettes describing each property.

This suggests that in general lexical cohesion will pick out a large unit on a certain topic, whereas in many cases one wants to separate out smaller units within that topic. The heuristic of grouping stuff that is lexically or semantically similar is opposed to the most common heuristic for segmentation in IE systems. A common approach in IE systems is to extract individual slots and fillers and then to *merge* information into frames, using the high level heuristic that one merges information unless it is contradictory. So if one has extracted a place followed by a date, then one combines them as two bits of information about one event, whereas if one has extracted two places in turn, then they probably describe different events. This heuristic involves seeing semantically cohesive words as an indicator of a *different* event rather than a continuation of the same topic. For instance, in the dialogue of Figure 3, the semantic cohesiveness of *3 days ago* and *yesterday* (as time expressions) or *injured* and *damaged* (as damage expressions) are indicative not of sentences 3–4 describing the same event as the rest, but rather of them describing a different (if related) event.

In this work I adopt the IE criterion of "combine unless contradictory", since it seems much more relevant to segmenting ads.[3] However, in line with work on text segmentation, segmentation is done as an early stage in processing (as also (Crowe1995)), rather than doing merging of information right at the end, as is more common in IE systems (Grishman and Sterling1993, Appelt and Israel1997).

[3]Unlike most other text segmentation work, note that the model of Beeferman et al.(1997) allows both attraction and repulsion between words, and therefore can generalize over both these intuitions, as appropriate to the domain.

## The problem domain and the information extraction system

The IE system runs over line ads from various Australian newspapers. Line ads have no structure beyond the presence of linebreaks. The first line is printed as bold, but may or may not be a real title. Linebreaks are sometimes clearly used with semantic significance, but in many other cases text continues across line breaks. The examples in this paper show the original line breaks. While the majority of ads advertise a single property, about 11% of ads advertise multiple properties. Since such ads commonly advertise four or more properties, bad performance on multi-property ads has a serious impact on the task-based evaluation of the IE system (on a per property, not a per advertisement basis). The nature of multiproperty ads varies considerably. Figures 1 and 2 shows three points in the space of possibilities.

How complicated can the hierarchical structure get? The most complex patterns usually occur in rental ads. In the example in Figure 1, one finds 4 levels of hierarchical structure, as shown in Figure 4. A third level of nesting can be observed in the first part of the ad in Figure 2(a). Most common is just two levels of structure: globally applicable information and information about individual houses. Even in such cases the scale of sharing can vary enormously. The only common information can be about the real estate agent (as in Figure 2(a) if we ignore temporarily the third level of nesting), or most of the information may be common as in Figure 2(b).

Details of the rest of the information extraction system are not central to this paper. Most of the system is handcrafted (this seemed the best way of meeting the deadlines and performance needs of the client). Standard learning techniques are used in some places (such as for categorizing ads). The system was designed to process only for sale ads, and so the results below will be restricted to this domain. It extracts individual facts with high accuracy (greater than 90% for all categories except suburb name (around 88%), which proves to be more difficult, in part because of the richness of abbreviations used).

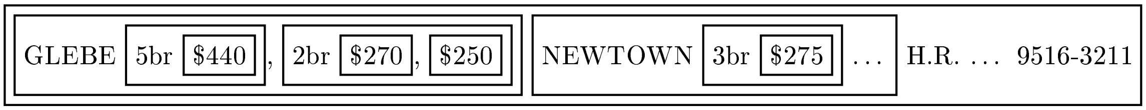


Figure 4. Levels of hierarchical structure in an advertisement

## Systems, results and evaluation

### The installed system

The text segmentation module in the installed system was also handcoded. It was developed quickly, and makes a number of quite crude assumptions. It does not consider segmentations between all tokens, but only at the ends of lines, after a semicolon (;) or before an asterisk (*). It does not allow arbitrary hierarchical structure but only two levels of hierarchy. These assumptions are adequate for over 95% of for sale ads. Information in the top level structure is inherited by individual houses in the information extraction process. The algorithm used is summarized in Figure 5. It segments off initial and final material that does not appear house-specific, and then starts a house specific ad. Lines are then added to a house-specific block until either (i) a line is read which overlaps too much in informational content with existing ads, or (ii) an ad looks complete and the next line looks like the beginning of a new house description. At this point a new house-specific block is begun. It is not straightforward what counts as compatible information, since for various reasons ads may have multiple specifications for a certain information type ("was $120,000, now for sale at $100,000"), and so non-contradictoriness was done on certain fairly reliable information (street address, number of bedrooms, inspection or auction info).[4] A house "has info and end" if it has at least two good indicators of a house specification (bedrooms, garages or square metres info; inspection or auction info; street address) and the last line ends with a period or separators like dashes or contains inspection info or a phone number (all common indicators of the end of an ad).

For the results reported here, the systems were run on 30 multiproperty advertisements which had not otherwise been used for development. For reporting results in terms of precision and recall, I score on each boundary where a nested segment begins or ends (I do not include the ends of the top level segment which must span the ad). This makes evaluation similar to previous work on segmentation, but rather different to standard evaluation in parsing (Collins1997), where a constituent span counts as a single item in assessing precision and recall.

[4] Segmentation was done before partial parsing, so phrasal information was not available to help disambiguate examples (such as the one above with two prices).

| Mistake category | Installed | Revised |
|---|---|---|
| Boundary off by one | 51.1% | 65.4% |
| Boundary off by two or more | 14.9% | 26.9% |
| Boundary missed | 31.9% | 7.7% |
| Boundary wrongly inserted | 2.1% | 0.0% |

Figure 6. An analysis of the segmentation errors of the installed system and the revised system. Note that percentages for the revised system are out of a smaller number of errors (though the percentage of one off errors is greater, their actual number is less).

On this basis, the installed system achieves a precision of 53.5% and a recall of 45.2%. These results are quite woeful. One might suspect that the overall IE system performs acceptably well only because multiproperty ads are quite rare. There is some truth in that, but a bit more is going on. Beeferman et al.(1997) deplore the use of precision and recall for evaluating segmentation algorithms, pointing out that it does not appropriately reward an algorithm which makes near misses. Each near miss scores one demerit point on both the precision and recall scales. One does better by missing a segment boundary altogether. Figure 6 shows that the majority of the errors of the system are in fact near misses. Beeferman et al.(1997) proceed to develop a metric based on the percentage of units correctly placed in the same segment that penalizes near mises less heavily. But is this largesse justified?

There are at least two reasons to not be entirely satisfied with the measure they propose. One is that it has limited dynamic range. They show that a random segmentation of the text scores 67% vs. 83% for their system (compare the figures for precision which give 17% vs. 56%). More seriously, this metric considers only the distance between the true boundary and the proposed boundary,

```
# line[1,...,m] have the (sub)line units text is divided into
# house[1,...] has the houses.
i = 1;
while (not house_specific(line[i]))
  i++;
k = m;
while (not house_specific(line[k]))
  k--;
m = 1;
for (j = i; j <= k; j++)
  if (not compatible_info(line[j], house[m]))
    m++;
  else if (has_info_and_end(house[m]) &&
           category(first(line[j])) in {money, suburb, asterisk})
    m++;
  house[m] = append(house[m], line[j]);
```

Figure 5. Installed text segmentation algorithm

without including the nature of the material over which there is disagreement. In general discourse, there is sometimes a graceful transition between topics, and the exact place to put a boundary is unclear, whereas in other cases there is a clear juncture. In the more prosaic world of information extraction, sometimes an incorrect boundary causes large mistakes, while in other cases the missegmented material is irrelevant. In Figure 7(a), the assigned boundaries get out of sync (the lack of a space in GREENVALLEY means the system fails to recognize it as a location), and the result is fairly catastrophic for the information extraction process because descriptions are being assigned to the wrong house. On the other hand, the example in Figure 7(b) shows a harmless slippage in boundary placement. Since the owner's mood is not part of the information extracted by the system, it makes no difference to system performance which of the two places the boundary is placed at. This confirms the general wisdom that a task-based evaluation is to be preferred to artificial evaluation metrics.

In Figure 8 I show IE results for the system overall (evaluated on other unseen data) versus on multiproperty ads alone, showing how bad segmentations significantly lower task-based performance.

## A revised system

Earlier testing of the installed system had already revealed its poor performance on multiproperty ads. The errors suggested that using a "merge unless contradictory" heuristic is insufficient for

```
INSPECT SAT'S: 2 - 2.45pm
=================================
GREENVALLEY 3 Norseman Close
---------------------------------
Brick home 3 bedroom, large
rumpus. Quiet cul-de-sac.
INSPECT: SAT'S 1.30 - 2.30pm
=================================
Hammondville 47/109 Stewart Ave
---------------------------------
2 bedroom villa. Complex includes
inground pool & tennis court.
INSPECT: SAT'S 2 - 3pm
=================================
MT PRITCHARD
10 & 12/347 Elizabeth Dr
---------------------------------
```

(a)

```
s.c. fire, air cond., country
kitchen with w.i. pantry. Large
verandah/pergola area. Garage
50' x 20'. $82,000.
-------------------------------
OWNER SAYS SELL
===============================
Adrian Davis (08) 8569 2120
WESTON RAINE & HORNE
96 Randell St. Mannum
```

(b)

Figure 7. Extracts from ads showing incorrect segment boundaries. Double lines show correct segment boundaries; single lines show chosen segment boundaries.

| | | | |
|---|---|---|---|
| Bedrooms | 46% | 92% | 91% |
| Price | 62% | 95% | 93% |
| Suburb | 63% | 88% | 90% |
| Street address | 50% | 93% | 90% |
| Agent Contact | 68% | 91% | 88% |
| Inspection | 71% | 94% | 94% |

Figure 8. Evaluation of information extraction performance of (i) the installed system on only multiproperty ads, (ii) the installed system on a general pool of ads (i.e., mainly one property per ad), and (iii) the revised system on only multi-property ads. (The evaluation gives accuracy results for 6 of the more common and important pieces of information in ads, but the system also attempts to extract many other types of information, such as auction details, garage and bathroom details, franchisor details, etc. The better performance of the revised system on suburbs may just be a fluke caused by the small evaluation sample, or could reflect the importance of the revised segmentation system on overall system performance.)

accurate boundary identification, and greater use had to be made of semantic types and typography that indicated a boundary, which were underutilized in the initial system. In particular, capitalization is often a clue to segmentation, but the initial system made no use of it. This suggests that previous text segmentation systems may have overly concentrated on lexical cohesion and repulsion rather than making sufficient use of a generalized notion of "discourse cues". By a generalized discourse cue, I mean anything (from typographic marks upwards) that is an indication of the beginning or end of a discourse segment. Hearst(1994) notes that the boundaries of her system are often off by a paragraph and suggests that incorporation of discourse cues might improve accuracy in such cases. I conjecture that the moderate performance levels of the algorithm of (Beeferman et al.1997) on finding exactly correct segment boundaries (recall and precision results in the fifties) suggests that their model of lexical attraction and repulsion also needs to be supplemented by a model sensitive to boundary features.

In the revised model, the earlier compatibility test is supplemented by replacing the "else if" clause of Figure 5 with a decision tree (Quinlan1986) trained on a different sample of hand-segmented ads. As shown in Figure 9, all but one of the questions the decision tree could choose from focus on generalized discourse cues in the

- Is the following unit all caps?
- Does the following unit start with a money amount?
- Does the following unit start with a suburb?
- Does the following unit start with a street address?
- Does the following unit start with a bedrooms specification?
- Does the following unit start with an asterisk?
- Does the next unit contain a phone number?
- Is it the beginning of the ad?
- Does the preceding unit end with a period?
- Does the preceding unit end with a semicolon?
- Does the preceding unit end with separators (dashes, etc.)?
- Does the preceding unit end with an inspection time?
- Does the preceding unit end with a phone number?
- Does the preceding unit contain a suburb?
- Does the preceding unit contain a money amount?
- Does the preceding house "have info" (same definition as before)?

Figure 9. Questions that the decision tree could ask

immediate context. This revised system gives a segment boundary precision of 70.7% and recall of 69% (a breakdown of the remaining errors appears in the righthand column of Figure 6). There are still a fair number of one-off errors, but the decision tree is particularly good at picking out good discourse cues for clear property junctures, and so most of the important errors are caught. Thus, the righthand column of Figure 8 shows that better use of generalized discourse cues increases task-based performance *more* than precision-recall evaluation, and indeed brings the level of performance of the system on multiproperty ads up to close to its performance on single property ads.

## Conclusions and Further Work

The main interest of the current work lies in the qualitative description of (i) what is required of text segmentation systems for use in practical applications, (ii) the basis on which text segmentation should proceed, and (iii) the need for task-based evaluation of text segmentations. Within a particular approach, I feel that there is still scope for improving quantitative results, and this is an area in which work is continuing.

## Bibliography


Douglas E. Appelt and David Israel. 1997. Building information extraction systems. Tutorial at ANLP-97.

Doug Beeferman, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Language Processing*, pages 35–46.

Mary Elaine Califf and Raymond J. Mooney. 1997. Relational learning of pattern-match rules for information extraction. In *ACL-97 Workshop in Natural Language Learning*.

Michael John Collins. 1997. Three generative, lexicalised models for statistical parsing. In *ACL 35/EACL 8*, pages 16–23.

Jeremy Crowe. 1995. Constraint-based event recognition for information extraction. In *ACL 33*, pages 296–298.

Ralph Grishman and John Sterling. 1993. Description of the Proteus system as used for MUC-5. In *Proceedings of MUC-5*. Morgan Kaufmann.

Barbara J. Grosz and Candance L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *ACL 32*, pages 9–16.

Hideki Kozima. 1993. Text segmentation based on similarity between words. In *ACL 31*, pages 286–288.

Diane J. Litman and Rebecca J. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In *ACL 33*, pages 108–115.

W. C. Mann and S. A. Thompson. 1987. Rhetorical structure theory: A theory of text organisation. Technical Report ISI/RS-87-190, Information Sciences Institute, Marina del Rey, CA.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.

Rebecca J. Passonneau and Diane J. Litman. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *ACL 31*, pages 148–155.

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.

Korin Richmond, Andrew Smith, and Einat Amitay. 1997. Detecting subject boundaries within text: A language independent statistical approach. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Language Processing*, pages 47–54.

Ellen Riloff. 1993. Automaatically constructing a dictionary for information extraction tasks. In *AAAI 11*, pages 811–816.

J. A. Rotondo. 1984. Clustering analysis of subject partitions of text. *Discourse Processes*, 7:69–88.

Stephen Soderland, D. Fisher, J. Aseltine, and Wendy Lehnert. 1995. Crystal: Inducing a conceptual dictionary. In *IJCAI 14*, pages 1314–1321.

Stephen Soderland. in press. Learning information extraction rules for semi-structured and free text. *Machine Learning*.