Extracting Data Records from Unstructured Biomedical Full Text

Donghui Feng **Gully Burns** **Eduard Hovy**
Information Sciences Institute
University of Southern California
Marina del Rey, CA, 90292
{donghui, burns, hovy}@isi.edu

Abstract

In this paper, we address the problem of extracting data records and their attributes from unstructured biomedical full text. There has been little effort reported on this in the research community. We argue that semantics is important for record extraction or finer-grained language processing tasks. We derive a data record template including semantic language models from unstructured text and represent them with a discourse level Conditional Random Fields (CRF) model. We evaluate the approach from the perspective of Information Extraction and achieve significant improvements on system performance compared with other baseline systems.

1 Introduction

The discovery and extraction of specific types of information, and its (re)structuring and storage into databases, are critical tasks for data mining, knowledge acquisition, and information integration from large corpora or heterogeneous resources (e.g., Muslea et al., 2001; Arasu and Garcia-Molina, 2003). For example, webpages of products on Amazon may contain a list of data records such as books, watches, and electronics. Automatic extraction of individual records will facilitate the access and management of data resources.

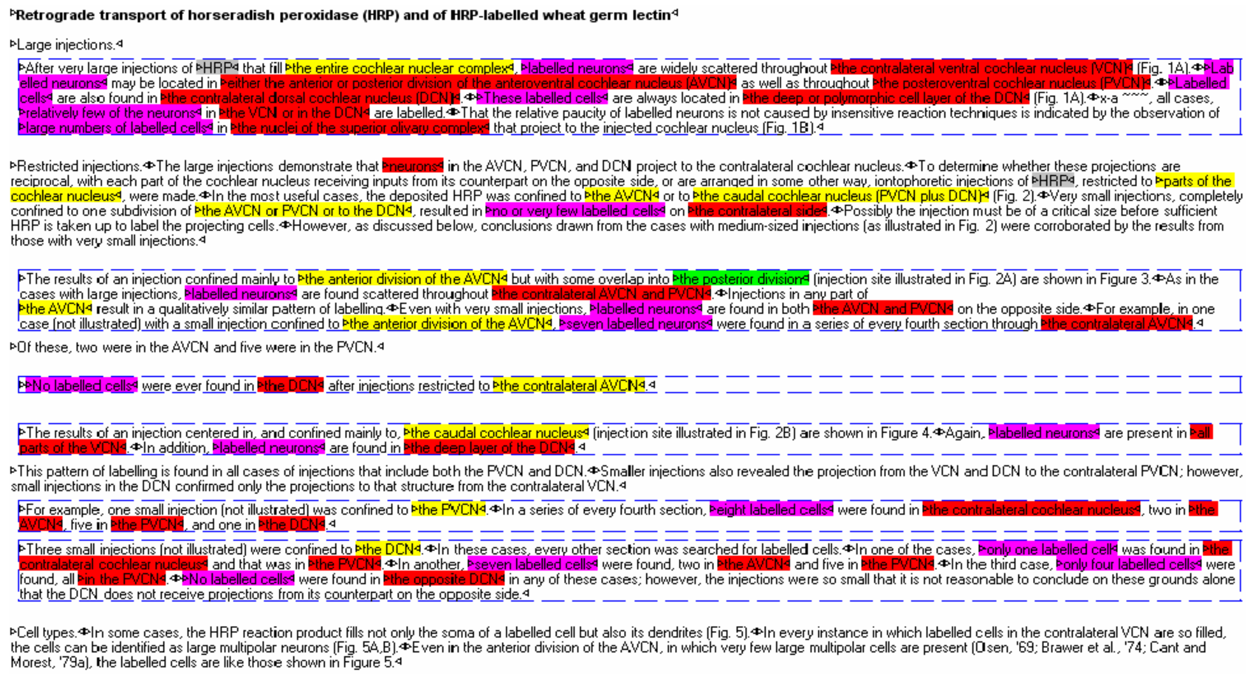
Most current approaches address this problem for structured or semi-structured text, for instance, from XML format files or lists and/or tabular data records on webpages (e.g., Liu et al., 2003; Zhu et al., 2006). The techniques applied rely strongly on the analysis of document structure derived from the webpage's html tags (e.g., the DOM tree model).

Regarding unstructured text, most Information Extraction (IE) work has focused on named entities (people, organizations, places, etc.). Such IE treats each extracted element as a separate record. Much less work has focused on the case where several related pieces of information have to be extracted to jointly comprise a single data record. In this work, it is usually assumed that there is only one record for each document (e.g., Kristjannson et al., 2004). Almost no work tries to extract multiple data records from a single document. Multiple data records can be scattered across the narrative in free text. The problem becomes much harder as there are no explicit boundaries between data records and no heavily indicative format features (like html tags) to utilize.

With the exponential increase of unstructured text resources (e.g., digitalized publications, papers and/or technical reports), knowledge needs have made it a necessity to explore this problem. For example, biomedical papers contain numerous experiments and findings. But the large volume and rate of publication have made it infeasible to read through the articles and manually identify data records and attributes.

We present a study to extract data records and attributes from the biomedical research literature. This is part of an effort to develop a Knowledge Base Management System to benefit neuroscience research. Specifically we are interested in knowledge of various aspects (attributes) of Tract-tracing Experiments (TTE) (data records) in neuroscience. The goal of TTE experiments is to chart the interconnectivity of the brain by injecting tracer chemicals into a region of the brain and identifying corresponding labeled regions where the tracer is

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 837–846, Prague, June 2007.



Retrograde transport of horseradish peroxidase (HRP) and of HRP-labelled wheat germ lectin

Large injections.

After very large injections of HRP that fill the entire cochlear nuclear complex, labelled neurons are widely scattered throughout the contralateral ventral cochlear nucleus (VCN) (Fig. 1A). Labelled neurons may be located in either the anterior or posterior division of the anteroventral cochlear nucleus (AVCN) as well as throughout the posteroventral cochlear nucleus (PVCN). Labelled cells are also found in the contralateral dorsal cochlear nucleus (DCN). These labelled cells are always located in the deep or polymorphic cell layer of the DCN (Fig. 1A). x-a ~~~, all cases, relatively few of the neurons in the VCN or in the DCN are labelled. That the relative paucity of labelled neurons is not caused by insensitive reaction techniques is indicated by the observation of large numbers of labelled cells in the nuclei of the superior olivary complex that project to the injected cochlear nucleus (Fig. 1B).

Restricted injections. The large injections demonstrate that neurons in the AVCN, PVCN, and DCN project to the contralateral cochlear nucleus. To determine whether these projections are reciprocal, with each part of the cochlear nucleus receiving inputs from its counterpart on the opposite side, or are arranged in some other way, iontophoretic injections of HRP, restricted to parts of the cochlear nucleus, were made. In the most useful cases, the deposited HRP was confined to the AVCN or to the caudal cochlear nucleus (PVCN plus DCN) (Fig. 2). Very small injections, completely confined to one subdivision of the AVCN or PVCN or to the DCN, resulted in no or very few labelled cells on the contralateral side. Possibly the injection must be of a critical size before sufficient HRP is taken up to label the projecting cells. However, as discussed below, conclusions drawn from the cases with medium-sized injections (as illustrated in Fig. 2) were corroborated by the results from those with very small injections.

The results of an injection confined mainly to the anterior division of the AVCN but with some overlap into the posterior division (injection site illustrated in Fig. 2A) are shown in Figure 3. As in the cases with large injections, labelled neurons are found scattered throughout the contralateral AVCN and PVCN. Injections in any part of the AVCN result in a qualitatively similar pattern of labelling. Even with very small injections, labelled neurons are found in both the AVCN and PVCN on the opposite side. For example, in one case (not illustrated) with a small injection confined to the anterior division of the AVCN, seven labelled neurons were found in a series of every fourth section through the contralateral AVCN.

Of these, two were in the AVCN and five were in the PVCN.

No labelled cells were ever found in the DCN after injections restricted to the contralateral AVCN.

The results of an injection centered in, and confined mainly to, the caudal cochlear nucleus (injection site illustrated in Fig. 2B) are shown in Figure 4. Again, labelled neurons are present in all parts of the VCN. In addition, labelled neurons are found in the deep layer of the DCN.

This pattern of labelling is found in all cases of injections that include both the PVCN and DCN. Smaller injections also revealed the projection from the VCN and DCN to the contralateral PVCN; however, small injections in the DCN confirmed only the projections to that structure from the contralateral VCN.

For example, one small injection (not illustrated) was confined to the PVCN. In a series of every fourth section, eight labelled cells were found in the contralateral cochlear nucleus, two in the AVCN, five in the PVCN, and one in the DCN.

Three small injections (not illustrated) were confined to the DCN. In these cases, every other section was searched for labelled cells. In one of the cases, only one labelled cell was found in the contralateral cochlear nucleus and that was in the PVCN. In another, seven labelled cells were found, two in the AVCN and five in the PVCN. In the third case, only four labelled cells were found, all in the PVCN. No labelled cells were found in the opposite DCN in any of these cases; however, the injections were so small that it is not reasonable to conclude on these grounds alone that the DCN does not receive projections from its counterpart on the opposite side.

Cell types. In some cases, the HRP reaction product fills not only the soma of a labelled cell but also its dendrites (Fig. 5). In every instance in which labelled cells in the contralateral VCN are so filled, the cells can be identified as large multipolar neurons (Fig. 5A,B). Even in the anterior division of the AVCN, in which very few large multipolar cells are present (Osen, '69; Brawer et al., '74; Cant and Morest, '79a), the labelled cells are like those shown in Figure 5.

Figure 1. An example of data records and attributes in a research article.

taken up and transported to (Burns et al., 2007).

To extract data records from the research literature, we need to solve two sub-problems: discovering individual attributes of records and grouping them into one or more individual records, each record representing one TTE experiment. Each attribute may contain a list of words or phrases and each record may contain a list of attributes.

Listing each sentence from top to bottom, we call the first problem the *Horizontal Problem* (HP) and the second the *Vertical Problem* (VP). Figure 1 provides an example of a TTE research article with colored fragments representing attributes and dashed frames representing data records. For instance, the third dashed frame represents one experiment record having three attributes with corresponding biological interpretations: "no labeled cells", "the DCN", and "the contralateral AVCN".

We view the HP and VP problems as two sequential labeling problems and describe our approach using two-level Conditional Random Fields (CRF) (Lafferty et al., 2001) models to extract data records and their attributes.

The HP problem (finding individual attribute values) is solved using a sentence-level CRF labeling model that integrates a rich set of linguistic features. For the VP problem, we apply a discourse-level CRF model to identify individual experiments (data records). This model utilizes deep semantic knowledge from the HP results (attribute labels within sentences) together with semantic language models and achieves significant improvements over baseline systems.

This paper mainly focuses on the VP problem, since linguistic features for the HP problem is the general IE topic of much past research (e.g., Peng and McCallum, 2004). We apply various feature combinations to learn the most suitable and indicative linguistic features.

The remainder of this paper is organized as follows: in the next section we discuss related work. Following that, we present the approach to extract data records in Section 3. We give extensive experimental evaluations in Section 4 and conclude in Section 5.

2 Related Work

As mentioned, data record extraction has been extensively studied for structured and semi-structured resources (e.g., Muslea et al., 2001; Arasu and Garcia-Molina, 2003; Liu et al., 2003; Zhu et al., 2006). Most of those approaches rely on the analysis of document structure (reflected in, for example, html tags), from which record templates are derived. However, this approach does not apply to unstructured text. The reason lies in the difficulty of representing a data record template in free text without formatting tags and integrating it into a learning system. We show how to address this problem by deriving data record templates through language analysis and representing them with a discourse level CRF model.

Given the problem of identifying one or more records in free text, it is natural to turn toward text segmentation. The Natural Language Processing (NLP) community has come up with various solutions towards topic-based text segmentation (e.g., Hearst, 1994; Choi, 2000; Malioutov and Barzilay, 2006). Most unsupervised text segmentation approaches work under optimization criteria to maximize the intra-segment similarity and minimize the inter-segment similarity based on word distribution statistics. However, this approach cannot be applied directly to data record extraction. A careful study of our corpus shows that data records share many words and phrases and are not distinguishable based on word similairties. In other words, different experiments (records) always belong to the same topic and there is no way to segment them using standard topic segmentation techniques (even if one views the problem as a finer-level segmentation than traditional text segmentation). In addition, most text segmentation approaches require a prespecified number of segments, which in our domain cannot be provided.

(Wick et al., 2006) report extracting database records by learning record field compatibility. However, in our case, the field compatibility is hard to distinguish even by a human expert. Cluster-based or pairwise field similarity measures do not apply to our corpora without complex knowledge reasoning. Most of Wick et al.'s data (faculty and student's homepages) contains one record.

In addition, as explained below, we have found that surface word statistics alone are not sufficient to derive data record templates for extraction. Some (limited) form of semantic understanding of text is necessary. We therefore first perform some sentence level extraction (following the HP problem) and then integrate semantic labels and semantic language model features into a discourse level CRF model to represent the template for extracting data records in the future.

Recently an increasing number of research efforts on text mining and IE have used CRF models (e.g., Peng and McCallum, 2004). The CRF model provides a compact way to integrate different types of features when sequential labeling is important. Recent work includes improved model variants (e.g., Jiao et al., 2006; Okanohara et al., 2006) and applications such as web data extraction (Pinto et al., 2003), scientific citation extraction (Peng and McCallum, 2004), and word alignment (Blunsom and Cohn, 2006). But none of them have used CRFs for discourse level data record extraction.

We use a CRF model to represent a data record template and integrate various knowledge as CRF features. Instead of traditional work on the sentence level, our focus here is on the discourse level. As this has not been carefully explored, we experiment with various selected features.

For the biomedical domain, our work will facilitate biomedical research by supporting the construction of Knowledge Base Management Systems (e.g., Stephan et al., 2001; Hahn et al., 2002; Burns and Cheng, 2006). Unlike the well-studied problem of relation extraction from biomedical text, our work focuses on grouping extracted attributes across sentences into meaningful data records. TTE experiment is only one of many experimental types in biology. Our work can be generalized to many different types of data records to facilitate biology research.

In the next section, we present our approach to extracting data records.

3 Extracting Data Records

Inspired by the idea of Noun Phrase (NP) chunking in a single sentence, we view the data records extraction problem as discourse chunking from a sequence of sentences using a sequential labeling CRF model.

3.1 Sequential Labeling Model: CRF

The CRF model addresses the problem of labeling sequential tokens while relaxing the strong independence assumptions of Hidden Markov Models (HMMs) and avoiding the presence of label bias from having few successor states. For each current state, we obtain the conditional probability of its output states given previously assigned values of input states. For most language processing tasks, this model is simply a linear-chain Markov Random Fields model.

In typical labeling processes using CRFs each token is viewed as a labeling unit. For our problem, we process each input document $D = (s_1, s_2, ..., s_n)$ as a sequence of individual sentences, with a corresponding labeling sequence of labels, $L = (l_1, l_2, ..., l_n)$, so that each sentence corresponds to only one label. In our problem, each data record corresponds to a distinct TTE experiment. Similar to NP chunking, we define three labels for sentences, "B_REC" (beginning of record), "I_REC" (inside record), and "O" (other). The default label "O" indicates that this sentence is beyond our concern.

The CRF model is trained to maximize the probability of $P(L \mid D)$, that is, given an input document D, we find the most probable labeling sequence L. The decision rule for this procedure is:

$$\hat{L} = \arg\max_{L} P(L \mid D) \quad (1)$$

A CRF model of the two sequences is characterized by a set of feature functions f_k and their corresponding weights λ_k. As in Markov fields, the conditional probability $P(L \mid D)$ can be computed using Equation 2.

$$P(L \mid D) = \frac{1}{Z_S} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k * f_k(l_{t-1}, l_t, D, t) \right) \quad (2)$$

where $f_k(l_{t-1}, l_t, D, t)$ is a feature function, representing either the state transition feature $f_k(l_{t-1}, l_t, D)$ or the feature of output state $f_k(l_t, D)$ given the input sequence. All these feature functions are user-defined boolean functions.

CRF works under the framework of supervised learning, which requires a pre-labeled training set to learn and optimize system parameters to maximize the probability or its log format. Equipped with this model, we investigate how to apply it and prepare features accordingly.

3.2 Feature Preparation

The CRF model provides a compact, unified framework to integrate features. However, unlike sentence-level processing, where features are very intuitive and circumscribed, it is not obvious what features are most indicative for our problem. We therefore explore three categories of features for discourse level chunking.

3.2.1 Semantic Attribute Labels

Most text segmentation approaches compute surface word similarity scores in given corpora without semantic analysis. However, in our case, data records have very similar characteristics and share most of the words. They are not distinguishable just from an analysis of surface word statistics. We have to understand the semantics before we can make decisions about data record extraction.

In our case, we care about the four types of attributes of each data record (one TTE experiment). Table 1 gives the definitions of the four attributes for each data record.

Name	Description
injectionLocation	the named brain region where the injection was made.
tracerChemical	the tracer chemical used.
labelingLocation	the region/location where the labeling was found.
labelingDescription	a description of labeling, including label density or label type.

Table 1. Attributes of data records (a TTE experiment).

To obtain this semantic attributes information of individual sentences (the HP problem), we first apply another sentence-level CRF model to label each sentence. We consider five categories of features based on language analysis. Table 2 shows the features for each category.

Name	Feature	Description
Lexicon Knowledge	TOPOGRAPHY	Is word topographic?
	BRAIN_REGION	Is word a region name?
	TRACER	Is word a tracer chemical?
	DENSITY	Is word a density term?
	LABELING_TYPE	Does word denote a labeling type?
Surface Word	Word	Current word
Context Window	CONT-INJ	If current word is within a window of injection context
Window Words	Prev-word	Previous word
	Next-word	Next word
Dependency Features	Root-form	Root form of the word if different
	Gov-verb	The governing verb
	Subject	The sentence subject
	Object	The sentence object

Table 2. The features for labeling words.

a. Lexicon knowledge. We used names of brain structures taken from brain atlases (Swanson, 2004), standard terms to denote neuro-anatomical topographical relationships (e.g., “rostral”), the name or abbreviation of the tracer chemical used (e.g., “PHAL”), and commonsense descriptions for descriptions of the labeling (e.g., “dense”, “light”).

b. Surface and window word. The current word and the words around are important indicators of the most probable label.

c. Context window. The TTE is a description of the inject-label-findings process. Whenever a word having a root form of “injection” or “deposit” appears, we generate a context window and all the words falling into this window are assigned a feature of “CONT-INJ”.

d. Dependency features. We apply a dependency parser MiniPar (Lin, 1998) to parse each sentence, and then derive four types of features from the parsing result. These features are (a) root form of every word, (b) the subject within the sentence, (c) the object within the sentence, and (d) the governing verbs.

The labeling system assigns a label for every token in each sentence. We achieved the best performance with an F-score of 0.79 (based on a precision of 0.80 and a recall of 0.78). This is not the focus of this paper. Please refer to our previous work (Burns et al., 2007) for details.

<SENT FILE="1995-360-213-ns.xml" INDEX= "63">
Regardless of the precise location of <tracerChemical> PHAL </tracerChemical> injection sites in <injectionLocation> the MEA </injectionLocation> , <labelingDescription> labeled axons </labelingDescription> followed the same basic routes .
</SENT>

Figure 2. An example of semantic attribute labels.

With the sentence-level understanding of each sentence, we obtain the semantic attribute labels for the data records. Figure 2 gives an example sentence with semantic attribute labels. Here <tracerChemical>, <labelingLocation>, and <labelingDescription> are recognized by the system, and the attribute names will be used as features for this sentence.

3.2.2 Semantic Language Model

Since text narratives might adhere to logical ways of expressing facts, language models for each sentence will also provide good features to extract data records. However, in biomedical research articles many of the technical words/phrases used in the narrative are repeated across experiments, making the surface word language model of little use in deriving generalized data record templates. Considering this, we replace in each sentence the labeled fragments with their attribute labels and then derive semantic language models from that format. By ‘semantic language model’ we therefore mean a combination of semantic labels and surface words.

For example, in the sentence shown in Figure 2, we have the semantic language model trigrams location-of-<tracerChemical>, sites-in-<injectionLocation>, and <labelingDescription>-followed-the. In addition, we also query WordNet for the root form of each word to generalize the semantic language models. This for example produces the semantic language model trigrams site-in-<injectionLocation> and <labelingDescription>-follow-the.

We believe the collected semantic language models represent an inherent structure of unstructured data records. By integrating them as features with a CRF model, we expect to represent data record templates and use the learned model to extract new data records.

However, it is not clear what semantic language models are most indicative and useful. A bag-of-words (language models) approach may bring much noise in. We show below a comparison of regular language models and semantic language models in evaluations.

3.2.3 Layout and Word Heuristics

The previous two categories of features come from the discovery of semantic components of sentences and their narrative form word analysis. When interviewing the neuroscience expert annotator, we learned that some layout and word level heuristics may also help to delineate individual data records. Table 3 gives the two types of heuristic features.

When a sentence contains heuristic words, it will be assigned to a word heuristic feature. If the sentence is at the boundary of a paragraph, it will be assigned a layout heuristic feature, namely the first or the last sentence in the paragraph.

Name	Feature	Description
EXP_B_WORD	INJECT CASE EXPERIMENT APPLICATION DEPOSIT PLACEMENT INTRODUCTION	Heuristic words for beginning of an experiment description
POS_IN_PARA	FIRST_IN_PARA LAST_IN_PARA	Position of the sentence in the paragraph

Table 3. The heuristic features.

4 Empirical Evaluation

To evaluate the effectiveness and performance of our technique, we conducted extensive experiments to measure the data record extraction approach.

4.1 Experimental Setup

We used the machine learning package MALLET (McCallum, 2002) to conduct the CRF model training and labeling.

We have obtained the digital publications of 9474 *Journal of Comparative Neurology (JCN)*[1] articles from 1982 to 2005. We have converted the PDF format into plain text, maintaining paragraph breaks (some errors still occur though). A simple heuristic based approach identifies semantic sections of the paper (e.g, Introduction, Results, Discussion). As most experimental descriptions appear in the Results section, we only process the Results section. A neuroscience expert manually annotated the data records in the Results section of 58 research articles. The total number of sentences in the Results section of the 58 files is 6630 (averaging 114.3 sentences per article).

	Training Set	Testing Set
Docs	39	19
Data Records	249	133

Table 4. Experiment configuration.

We randomly divided this material into training and testing sets under a 2:1 ratio, giving 39 documents in the training set and 19 in the testing set. Table 4 gives the numbers of documents and data records in the training and the testing set.

4.2 Evaluation Metrics

To evaluate data record extraction, we notice it is not fair to strictly evaluate the boundaries of data records because this does not penalize the near-miss and false positive of data records in a reasonable way; sentences near a boundary that contain no relevant record information can be included or omitted without affecting the results. Hence the standard P_k (Beeferman et al., 1997) and *WinDiff* (Pevzner and Hearst, 2002) measures for text segmentation are not so suitable for our task.

As we are concerned with the usefulness of knowledge in extracted data records, we instead evaluate from the perspective of IE. We measure system performance on the quality of the extracted data records. For each extracted data record, it will be aligned to one of the data records in the gold standard using the "dominance rule" (if the data record can be aligned to multiple records in the gold standard, it will be aligned to the one with highest overlap). Then we evaluate the precision, recall, and F1 scores of extracted units of the data record. The units are the attributes in data records.

$$precision = \frac{\#\, of \text{ correct units}}{\#\text{ of the extracted units by the system}} \quad (3)$$

$$recall = \frac{\#\, of \text{ correct units}}{\#\text{ of the units in the gold standard}} \quad (4)$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad (5)$$

These measures provide an indication of the completeness and correctness of each extracted record (experiment). We also measure the number of distinct records extracted, compared with the gold standard as appearing in the document.

4.3 Experiment Results

To fully compare the effectiveness of our semantic analysis functionality, we evaluated system performance for all the following systems:

TextTiling (TT): To compare with text segmentation techniques, we use TextTiling (Hearst, 1994) with default parameters as the first baseline system.

Random Guess (RG): In order to demonstrate the data balance of all the possible labels in the testing set, we also use another baseline system with random decisions for each sentence.

[1] http://www3.interscience.wiley.com/cgi-bin/jhome/31248

Domain Heuristics (DH): In a regular TTE experiment, only one tracer chemical will typically be used. Given this heuristic, we assume each data record contains one tracer chemical. In this system, we first locate sentences with identified trace chemicals, and then we greedily expand backward and forward until another new tracer chemical appears or no other attribute is included.

Surface Text (ST): To measure the effectiveness of the semantic analysis (attribute labels and semantic language models), the ST system utilizes only standard surface word language models and heuristic features.

Semantic Analysis (SEM): The SEM system uses all the semantic features available (including identified attributes and semantic language models) and two heuristic features.

Table 5 shows the final performance of these different systems. The second column provides the numbers of extracted data records. In this task, a larger number does not necessarily mean a better system, as a system might produce too many false positives. The remaining three columns represent the precision, recall, and F1 scores, averaged over all data records. With our approach, the system performance is significantly improved compared with other systems. System TT fails in this task as it only outputs the full document as one single record.

	# of Records	Prec.	Rec.	F1
TT	19	0.3861	1.0	0.5571
RG	758	0.6331	0.0913	0.1595
DH	162	0.6703	0.4902	0.5663
ST	82	0.8182	0.8339	0.8260
SEM	72	**0.8505**	**0.9258**	**0.8865**

Table 5. System performance.

To investigate how plain text language models and semantic language models affect system performance, we also experimented with all the language models. Table 6 shows comparisons of three types of language models. Systems with semantic analysis always work better than those with only surface text analysis. Without semantic analysis, unigram features work better than bigram and trigram features. This matches our intuition: without generalizing to semantic language models, higher order language models will be relatively sparse and contain much noise. However, when taking into account the semantic features, we found that bigram and trigram semantic language model features outperformed unigrams. They are especially important in boosting the recall scores as they capture more generalized information when derived.

	Unigram (%)	Bigram (%)	Trigram (%)
	Prec/Rec/F1	Prec/Rec/F1	Prec/Rec/F1
ST	**81.8/83.4/82.6**	69.1/88.4/77.6	57.9/88.8/70.1
SEM	85.1/86.6/85.6	**85.1/92.6/88.7**	82.2/92.7/87.1

Table 6. Language model comparisons.

As an example, Table 7 gives a list of high quality bigram semantic language models ranked by their information gains based on the training data.

through_<labelingLocation>	rat_no
<labelingDescription>_be	of_<tracerChemical>
<labelingLocation>_(	<tracerChemical>_be
<tracerChemical>_injection	be_inject
into_<injectionLocation>	be_center
<labelingDescription>_from	inject_with
<tracerChemical>_in	injection_of
in_<labelingLocation>	in_experiment

Table 7. An example list of top-ranked bigrams.

The main difficulty for data record extraction from unstructured text lies in deriving and representing a template for future extraction. We actually take advantage of CRF and represent the template with a CRF model.

Each data record is measured with precision, recall, and F1 scores. Figure 3 depicts the distribution of extracted data records according to these measures in the best system.

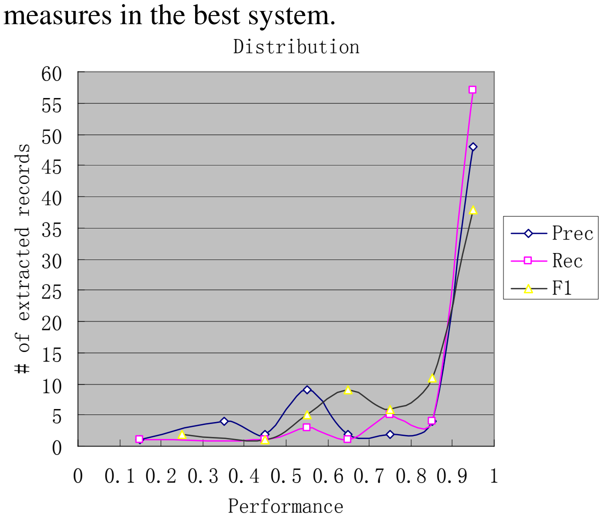


Figure 3. Data records performance distribution.

The results are encouraging, especially given the complexity and flexibility of data record descriptions in the unstructured text. In Figure 3, Axis X represents the value interval for precision, recall, and F1, and Axis Y represents the number of extracted records with their corresponding values. For example, 57 records have recall scores falling into [0.9, 1.0].

Figure 4 gives an example alignment between system result and the gold standard. Each record is represented by a range of sentences. The numbers following each record in the system result are individual data record's precision and recall scores.

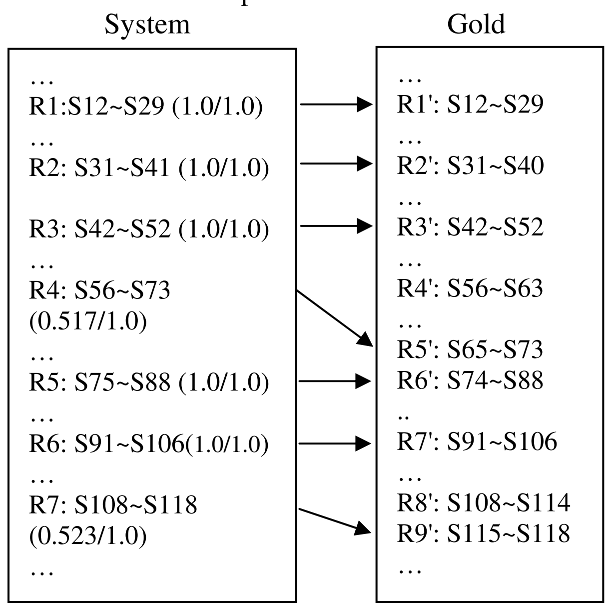


Figure 4. An example of record extraction in one doc.

This is a real example from the testing set. For records R1, R3, and R6, the system can extract the exact sentences contained. For record R2 and R5, although they do not exactly match at the sentence level, the extracted record contains the entire required set of attributes as in the gold standard.

4.4 Error Analysis and Discussion

When we investigated the errors, we found that sometimes the extracted data records combined two or more smaller gold standard records, or vice versa. As shown in Figure 4, extracted records R4 and R7 are both combinations of records in the gold standard. This is partially due to the granularity definition problem. Authors may mention several approaches/symptoms to one type of experiment for a single purpose. In this case, it is almost infeasible to have annotators strictly agree on granularity and thus to teach the system to acquire this knowledge. For example, in the gold standard, the annotator annotated three successive sentences as three separate records but the system output those as only one data record. In this extreme case, it is too hard to expect the system to perform well.

In our approach, the semantic attribute labels and semantic language models require the result of the initial sentence-level labeling, which has an F-score of 0.79. The error may propagate into the data record extraction procedure and lower overall system performance.

In our current experiments, we also assume all the attributes within one segment belong to one record. However, the situation of embedded data records will make this problem harder. For example, authors sometimes compare the current experiment with other approaches in referenced papers. In this case, those attributes should be excluded from the records. We need to invent rules or constraints to filter them out. When such reference occurs at experiment boundaries, it brings higher risk for correct results.

It is a very hard problem to extract from unstructured text neat structured records. The annotators sometimes employ background knowledge or reasoning when performing manual extraction; such knowledge cannot today be easily modeled and integrated into learning systems.

In our study, we also compared some feature selection approaches. Similar to (Yang and Pedersen, 1997), we tried Feature Instance Frequency, Mutual Information, Information Gain, and CHI-square test. But we eventually found that the system including all the features worked best, and with all the other configurations unchanged, feature instance frequency worked at almost the same level as other complex measures such as mutual information and information gain.

5 Conclusion and Future Work

In this paper, we explored the problem of extracting data records from unstructured text. The lack of structure makes it difficult to derive meaningful objects and their values without resorting to deeper language analysis techniques. We derived indicative linguistic features to represent data record templates in free text, using a two-pass approach in which the second pass used the IE labels derived from the first to compose attributes into coherent data records. We evaluated the results from an IE perspective and reported potential problems of error generation.

For the future, we plan to explore additional feature types and feature selection strategies to determine what is "good" for unstructured record templates to improve our results. More effort will also be put into the sentence-level analysis to reduce error propagations. In addition, ontology based knowledge inference strategies might be useful to validate attributes in single record and in turn help data record extraction. The last thing under our direction is to explore new models if applicable.

We hope this thought-provoking problem will attract more attention from the community. In the future, we plan to make our corpus available to the community. The solution to this problem will highly affect the access of knowledge in large scale unstructured text corpora.

Acknowledgements

The work was supported in part by an ISI seed funding, and in part by a grant from the National Library of Medicine (RO1 LM07061). The authors want to thank Feng Pan for his helpful suggestions with the manuscript. We would also like to thank the anonymous reviewers for their valuable comments.

References

Arasu, A., and Garcia-Molina, H. 2003. Extracting structured data from web pages. In *Proc. of SIMOD-2003*.

Beeferman, D., Berger, A., and Lafferty, J. 1997. Text segmentation using exponential models. In *Proc. of EMNLP-1997*.

Blunsom, P. and Cohn, T. 2006. Discriminative word alignment with conditional random fields. In *Proc. of ACL-2006*.

Brazma, A., et al., 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet,* 29(4): p. 365-71.

Burns, G.A. and Cheng, W.-C. 2006. Tools for knowledge acquisition within the NeuroScholar system and their application to anatomical tract-tracing data. In *Journal of Biomedical Discovery and Collaboration*.

Burns, G., Feng, D., and Hovy, E.H. 2007. Intelligent Approaches to Mining the Primary Research Literature: Techniques, Systems, and Examples. Book Chapter in *Computational Intelligence in Bioinformatics,* Springer-Verlag, Germany**.**

Choi, F. Y. Y. 2000. Advances in domain independent linear text segmentation. In *Proc. of NAACL-2000*.

Hahn, U., Romacher, M., and Schulz, S. 2002. Creating knowledge repositories from biomedical reports the MEDSYNDIKATE text mining system. In *Proc. of PSB-2002*.

Hearst, M. 1994. Multi-paragraph segmentation of expository text. In *Proc. of ACL-1994*.

Jiao, F., Wang, S., Lee, C., Greiner, R., and Schuurmans, D. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proc. of ACL-2006*.

Kristjannson, T., Culotta, A. Viola, P., and McCallum, 2004. A. Interactive information extraction with constrained conditional random fields. In *Proc. of AAAI-2004*.

Lafferty, J., McCallum, A. and Pereira, F. 2001 Conditional Random Fields: probabilistic models for segmenting and labeling Sequence Data. In *Proc. of ICML-2001*.

Lin, D. 1998. Dependency-based evaluation of MINIPAR. In *Proc. of Workshop on the Evaluation of Parsing Systems*.

Liu, B., Grossman, R., and Zhai, Y. 2003. Mining data records in web pages. In *Proc. of SIGKDD-2003*.

Malioutov, I. and Barzilay, R. 2006. Minimum cut model for spoken lecture segmentation. In *Proc. of ACL-2006.*

McCallum, A.K. 2002. *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Muslea, I., Minton, S., and Knoblock, C.A. 2001. Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems* 4:93-114.

Okanohara, D., Miyao, Y., Tsuruoka, Y., and Tsujii, J. 2006. Improving the scalability of semi-markov conditional random fields for named entity recognition**.** In *Proc. of ACL-2006*.

Peng, F. and McCallum, A. 2004. Accurate information extraction from research papers using conditional random fields. In *Proc. of HLT-NAACL-2004*.

Pevzner, L., and Hearst, M. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*.

Pinto, D., A. McCallum, X. Wei, and W.B. Croft. 2003. Table Extraction Using Conditional Random Fields. In *Proc. of SIGIR-2003*.

Stephan, K.E. et al., 2001. Advanced database methodology for the Collation of Connectivity data on the Macaque brain (CoCoMac). *Philos Trans R Soc Lond B Biol Sci*, 356(1412).

Swanson, L.W. 2004. *Brain Maps: Structure of the Rat Brain*. 3rd edition, Elsevier Academic Press.

Wick, M., Culotta, A., and McCallum, A. 2006. Learning field compatibilities to extract database records from unstructured text. In *Proc. of EMNLP-2006.*

Yang, Y., and Pedersen, J. 1997. A comparative study on feature selection in text categorization. In *Proc. of ICML-1997*, pp. 412-420.

Zhu, J., Nie, Z., Wen, J., Zhang, B., and Ma, W. 2006. Simultaneous record detection and attribute labeling in web data extraction. In *Proc. of KDD-2006*.