A Comparative Study
# of Information Extraction Strategies

Ronen Feldman, Yonatan Aumann, Michal Finkelstein-Landau,
Eyal Hurvitz, Yizhar Regev, Ariel Yaroshevich

ClearForest Ltd.
Or Yehuda, Israel
Tel. 972 3 7350000
{ronen, yonatan, michal, eyal, ryizhar, ariel}@clearforest.com

Abstract. The availability of online text documents exposes readers to a vast amount of potentially valuable knowledge buried therein. The sheer scale of material has created the pressing need for automated methods of discovering relevant information without having to read it all. Hence the growing interest in recent years in Text Mining.
A common approach to Text Mining is Information Extraction (IE), extracting specific types (or templates) of information from a document collection. Although many works on IE have been published, researchers have not paid much attention to evaluate the contribution of syntactic and semantic analysis using Natural Language Processing (NLP) techniques to the quality of IE results.
In this work we try to quantify the contribution of NLP techniques, by comparing three strategies for IE: naïve co-occurrence, ordered co-occurrence, and the structure-driven method – a rule-based strategy that relies on syntactic analysis followed by the extraction of suitable semantic templates. We use the three strategies for the extraction of two templates from financial news stories. We show that the structure-driven strategy provides significantly better precision results than the two other strategies (80-90% for the structure-driven compared with about only 60% for the co-occurrence and ordered co-occurrence). These results indicate that a syntactical and semantic analysis is necessary if one wishes to obtain high accuracy.

Keywords: Information Extraction, Natural Language Processing

1. INTRODUCTION

The need for automated methods to extract information from online text documents is constantly growing in this age of information overload. Since most of the information available in digital format is unstructured, there is a growing interest in Text Mining that focuses on extracting data from textual sources.

One of the most common processes of Text Mining is known as Information Extraction (IE).

In Information Extraction, key concepts (facts or events concerning entities or relationships between entities discussed in the text) are defined in advance and then the text is searched for concrete evidence for the existence of such concepts. For example, in financial news documents we may be interested in information about acquisitions of a certain company by another company. Such information may be typically given by a sentence such as:

ABC Inc, the leading manufacturer of electronic toys, has successfully completed the acquisition of DFG Corporation.

The above would be converted into a fact, or an instance of the template:

Acquisition:	Company1	Company2
	ABC Inc.	DFG Corporation

Table 1. Acquisition template

Thus, structured information is created from the unstructured text.

Another example is what we will call Person-Left-Position: information about the fact that a certain employee in a company left the company (willingly or unwillingly).

An example of a sentence delivering this information is:

Andrx Group (Nasdaq:ADRX) today announced that Chih-Ming Chen, Ph.D. has resigned from his position as the Company's Chief Scientific Officer.

The corresponding template is:

Person-Left-Position:	Company	Person	Position
	Andrx Group	Chih-Ming Chen	Chief Scientific Officer

Table 2. Person-Left-Position template

There are several different algorithms and methods to perform Information Extraction. These are based on various levels of semantic and syntactic analysis of the text.

Existing IE systems include systems based on hand-crafted rules that “understand” the text and manage the filling of the template slots, as ([1], [4]), as well as trainable systems (WAVE [2], CRYSTAL ([5],[6]).

Trainable systems have the advantage over hand-crafted systems that they can be extended more easily and require less domain knowledge. However, the performance of the trainable systems is usually not as good as the hand-crafted ones (precision and recall-wise).

In this paper we compare the performance of three strategies for Information Extraction. We show a general method for performing semantic and syntactic analysis of the text that enables constructing of "structure-driven rules" that achieve high levels of precision (80%-90%).

We compare it to two other strategies: Co-occurrence strategy and Ordered Co-occurrence strategy. The co-occurrence strategy is much simpler than the structure-driven strategy, seeking only the existence of relevant keywords in the text, without reference to their syntactic or semantic role therein. The Ordered Co-occurrence strategy is similar to the Co-occurrence strategy, but here constraints regarding the position of the keyword within the sentence, relatively to the entities involved in the template are applied. We show that, while these two strategies allow rapid constructing of rules, the precision of such rules is consistently relatively low (50%-60%). That is: although the structure-driven rules require more labor, the precision results clearly justify the additional work.

The remainder of this paper is organized as follows. In Section 2 we describe the three strategies in details. In Section 3 we describe the experimental evaluation of the three strategies in the DIAL language. In Section 4 we describe the comparison performed and its results. We discuss the results in Section 5.

2. INFORMATION EXTRACTION STRATEGIES

An advanced task in Information Extraction systems involves filling up predefined templates after identifying semantic relationships between different entities in the text.

In this section we will describe a hierarchy of three strategies, evaluated and compared in this study, for extracting semantic relationships between entities in an Information Extraction system. The hierarchy ranges from a simple co-occurrence method to a sophisticated rule-based method for extracting information from a single sentence.

All strategies are based on an underlying system for entity recognition, namely a system that extracts proper names and classifies them according to a predefined set of categories, such as Company, Person, Location and so forth. The identified entities are utilized by the different methods as-is without any further analysis.

For illustration purposes, we will use a binary relation between companies called ACQUIRED:

ACQUIRED(Company1, Company2) means that an acquisition event took place between Company1 and Company2, namely either Company1 acquired Company2 or the opposite.

2.1 The Co-occurrence Strategy

This method follows the simple definition of the term co-occurrence: “an event or situation that happens at the same time as or in connection with another” (http://www.dictionary.com). It seeks only the existence of the relevant entities and keywords in the same sentence.

This method is implemented by a simple pattern matching mechanism without referring to any syntactic or semantic role of the searched entities and keywords.

For identifying the ACQUIRED relationship, this strategy searches for sentences that contain the following elements:

- Two different companies: identified at an earlier stage of entity recognition, as described above
- Acquisition keyword: a keyword taken from a lexicon of acquisition nouns and verbs; e.g. acquire, bought, acquisition and so on.

The following sentences were extracted as candidates for the ACQUIRED relationship by the Co-occurrence strategy.

The first one is correct; the second is incorrect. Co-occurrence elements (companies and acquisition keywords) are bolded.

- *Recently,* ***Sovereign*** *entered into a definitive agreement with* ***Main Street Bancorp, Inc.*** *(“Main Street”) for* ***Sovereign*** *to* ***acquire Main Street****.*
- ***Ask Jeeves*** *deploys its solutions on* ***Ask Jeeves*** *at* ***Ask.com****,* ***Ask Jeeves for Kids*** *at* ***AJKids.com****,* ***DirectHit.com*** *and* ***Jeeves Tours****, to help companies target and* ***acquire*** *qualified prospects online and to provide consumers with real-time access to information, products and services.*

2.2 The Ordered Co-occurrence Strategy

This method is an enhancement of the naïve co-occurrence method. We choose to enhance the simple co-occurrence results by adding order constraints on the matched pattern. Such constraints are intended to heuristically preclude the extraction of syntactically invalid or semantically unreasonable events.

The simple co-occurrence strategy might extract sentences where the searched elements by no means form a valid structure for correct semantic relationships. For example, if the keyword searched for is a transitive verb, it should be located between the entities, neither precede them nor follow them.

Within the ACQUIRED relationship, forcing the keyword “acquire” to separate one company from the other helps in eliminating trivial precision errors like in the following sentence, where the transitive verb precedes both companies:

The building was vacant at the time it was ***purchased*** *and is now 100% leased to* ***Deltek Systems*** *and* ***Perot Systems*** *70,524 square feet executed in late July of 2001).*

Defining the appropriate constraints requires a shallow linguistic understanding of the domain in order to determine the appropriate order between the searched elements.

2.3 The Structure-Driven Rule-Based Strategy

This strategy is based on noun phrase and verb phrase identification augmented by linguistic and semantic constraints.

In this strategy, the extraction of the predefined semantic relationships is performed in the means of deep syntactic and semantic analysis of the sentences. Naturally, this method involves more human effort, but we will show that it consistently achieves higher precision rates.

For example, for the ACQUIRED relationship we search for a Subject-Verb-Object structure, requiring the Subject and Object to be companies and the Verb to be tensed where its head belongs to the Acquisition lexicon (e.g. *acquire*, *purchase*). The constraints require, for example, different Subject and Object (i.e. two different companies - a semantic constraint) and verb-preposition agreement (syntactic constraint).

As indicated by this example, this method requires a skilled developer and entails a fairly elaborate development effort. The advantage, as will be discussed later in this document, is that its qualitative results are by far better than the two simpler methods.

The implementation of the Structure-Driven processing is based on a general multi-level NLP system. We give here a brief description of its different layers:

Layer 0 - POS (Part of Speech) Tagger: Assigning POS tags (noun, proper noun, verb, adjective, adverb, preposition, and so on.) to each word.

Layer 1 - Noun Phrase and Verb Phrase Grouper: Grouping together the head noun with its left modifiers (for example: "*massive payment agreement*") and, for verbs, chunking a main verb with its auxiliaries, like in "*has been acquired*" or "*is already being incorporated*".

Layer 2 - Verb and Noun Pattern Extractor: Extracting larger verb and noun phrases, on the basis of semantic requirements. Examples: "*said Monday it has acquired*" and "*announced plans to acquire*".

In general, this mechanism matches verbs and nouns with their complements, as specified in their sub-categorization properties. This level is semantically-oriented: it keeps track of the semantic features of a pattern, as expressed by various elements such as adverbs, tense and voice of the verb group and certain syntactic structures. This way, the system can identify complex patterns that still express a basic relation given by the rightmost element of the pattern. For example, in "*SignalSoft has expanded its application portfolio with the acquisition of mobilePosition(R)*", "*has expanded its application portfolio with the acquisition*" is a Verb Pattern based on the keyword "*acquisition*", that is used to extract acquirer-acquired relations.

Layer 3 - Named Entity Recognizer: recognition of companies, persons, products, and so forth.

Layer 4 - Nominal Expression Extractor: Matching nominal phrases that contain entities as arguments, such as "*Microsoft's acquisition of Visio*", or "*The acquisition by Microsoft of Visio*".

Layer 5 - Template ("Event") Extractor: Rule-based extraction of patterns at a full sentence or phrase level.

For example, the full sentence “*Microsoft announced Monday it has acquired Visio*” is matched using the Verb Pattern of Layer 2 “*announced Monday it has acquired*”. This layer uses a lexicon of keywords, nouns and verbs that are relevant to the specific template. (For example, in the case of the Acquisition template, verbs such as “*acquire*”, “*buy*”, “*bid*”). This layer includes extraction of other elements that are needed to shallow parse sentences and additional information regarding a template (such as adverbial phrases, appositive clauses, dates, and so forth.).

3. Implementation in the DIAL Extraction Language

In this section we will briefly describe the framework used for building our IE system, a rule-based general IE language developed at ClearForest (DIAL).

DIAL is a declarative, rule-based language, designed specifically for IE. The complete syntax of DIAL is beyond the scope of this paper. In the next item we present of the key elements relevant to this work. Further details and examples are presented in [3].

DIAL enables the user to implement separately the different operations required for performing IE: tokenization, zoning (recognizing paragraph and sentence limit), and morphological and lexical processing, parsing and domain semantics. DIAL has built-in modules that perform the general tasks of tokenization and part-of-speech tagging. In addition, we have developed a general library of rules that perform Noun Phrase and Verb Phrase grouping and separate libraries for recognizing relevant Entities, such as *companies* or *persons*.

3.1 Survey of DIAL's Basic Elements

As stated above, DIAL is a rule-based language. DIAL “program” is phrased as a logic program - a Rule Book.

A Rule Book, , is a conjunction of Definite clauses (“rules”) $C_i : H_i \leftarrow B_i$, where C_i is a clause label, H_i (“the head”) is a literal and $B_i = (B_{i1},B_{i2} \ldots) = (P_i,N_i)$ (the clause's body), where $P_i = (p_{ij})$ is a series of Pattern Matching Elements and $N_i = \{n_{ij}\}$ is a set of constraints operating on P_i.

The clause $C_i : H_i \leftarrow B_i$ represents the assertion that H_i is implied (or, in our context, that an instance of H_i is defined) by the conjunction of the literals in P_i while satisfying all the constraints in N_i.

Typically, the $H_i$is the template (event) sought by the Information Extraction process (such as Acquisition or Person-Left-Position). The practical meaning of the above formal definition is that whenever the series of pattern matching elements P_i is found in the text and the constraints set N_i is fulfilled , deduce that the template H_i occurs in that text fragment.

A Pattern Matching Element p_{ij} may be:

- An explicit token (String) found in the text - e.g. *"announces"*
- A word class element: a phrase from a predefined set of phrases that share a common semantic function. Example: the word class *wcResignation* includes the words: *"resignation"*, *"retirement"* and "*departure*".
- A predicate call - e.g. Company(C)

See [3] for a more complete list of DIAL elements.

A constraint n_{ij} may be used for carrying out on-the-fly Boolean checks on relevant segment of texts matched by the pattern matching elements. A constraint is typically implemented by using a suitable Boolean function, for example: InWC, which returns TRUE if the tested text segment is a member of the tested word class.

For example, verify(InWC(*P*, @wcAnnounce)) means that the *P* pattern matching element must be a member of the word class wcAnnounce.

3.2 DIAL Rules - an Example

Below we give an example of a rather simple DIAL rule for extracting a common Person-Left-Position template:

```
PersonLeftPosition(Person_Name, Position, Company_Name) :-
Company(Company_Name)
Verb_Group(V_Stem,V_Tense,V_Modifiers)

Noun_Group(N_Determiner,N_Head,N_Stem,N_Modifiers)
“of”
Person(Person_Name)
[ “as” ]
wcCompanyPositions
verify(InWC(V_Stem,@wcAnnounce))
verify(InWC(N_Stem,@wcResignation)) ;
```

The rule above corresponds to a common pattern in financial news announcing resignation or retirement, as in: "*International Isotopes Inc Announces the Resignation of Dr. David Camp As President and CEO*".

The meaning of the code above is as follows: Extract a *Person-Left-Position* template from this text segment if a *Company* was identified, followed by a Verb Group whose stem is included in the *wcAnnounce* word class (that includes verb such as "*announce*" or "*report*"), followed by a Noun Group (that may include a determiner such as "*the*") whose head is a member of the word class *wcResignation* (This word class includes the terms "*resignation*", "*retirement*" and "*departure*"), followed by the word "*of*", followed by a *person* name, followed by the optional word "*as*" and a term from *wcCompanyPosition*, a word class that includes common positions of executives such as "*President*", "*CEO*", "*CFO*" and so forth.

The Company and Person predicates are implemented in a separate module that is executed before the *Person-Left-Position* module.

4. Experimental Evaluation

In order to test the three strategies we have conducted two separate experiments. In each experiment we tested the results of the three strategies for the extraction of one concept (template) – in one experiment the *Acquisition* template was extracted, and in the second experiment – the *Person-Left-Position* template.

4.1 The Data Source

For each experiment, we created a news article collection by downloading documents from the *NewsAlert* site (http://www.newsalert.com) using a suitable set of keywords:

For the *Acquisition* template, the keyword set included terms such as "acquisition", "acquire", "buy", "bid" and "purchase". The collection included 500 document published in September 2001.

For the *Person-Left-Position,* the keyword set included terms such "resign", "retire","resignation", "fire" and "step down". The collection included 1725 documents published in August-September 2001. (The number of required documents for this template was bigger than the number required for the *Acquisition* template, because this template is less frequent).

The *NewsAlert* site aggregates news document from a number of sources, including, among others, *Reuters*, *PRNewsWire* and *BusinessWire.*

4.2 Evaluating the Different Strategies

It is important to note that the set of structure-driven rules were written prior to downloading the test collection, using the DIAL NLP libraries described in section 2.3 above. These rules were written based on a small set of financial news documents we had previously downloaded. We have also created sets of rules implementing the two other strategies.

For each of the two templates, we executed separately the rules written according to each of the three strategies using the ClearStudio environment developed at ClearForest. The ClearStudio environment creates as a result a file of all the instances found and enables viewing the location within the original document from which the instance was extracted and to classifying that instance (for example, as correct or incorrect). See Figure 1 below.

In the *Acquisition* template, all rules were required to extract the two companies involved in the Acquisition relationship. For the purpose of the experiment, we ignore modalities, so that an extracted Acquisition event could be either an actual, possible, pending or even a cancelled acquisition. In the *Person-Left-Position* template, an instance was extracted if it included the person name and the company from which she or he retired, or, the person name and the position she or he had held.

For each of the two templates, we executed separately the rules written according to each of the three approaches using the ClearStudio environment developed at ClearForest. The ClearStudio environment creates as a result a file of all the instances found and then, to view the location within the original document from which the instance was extracted and to classify that instance (For example, as correct or incorrect). See Figure 1 below.

4.3 The Results

The results for the two templates are given in tables 3 and 4 below.

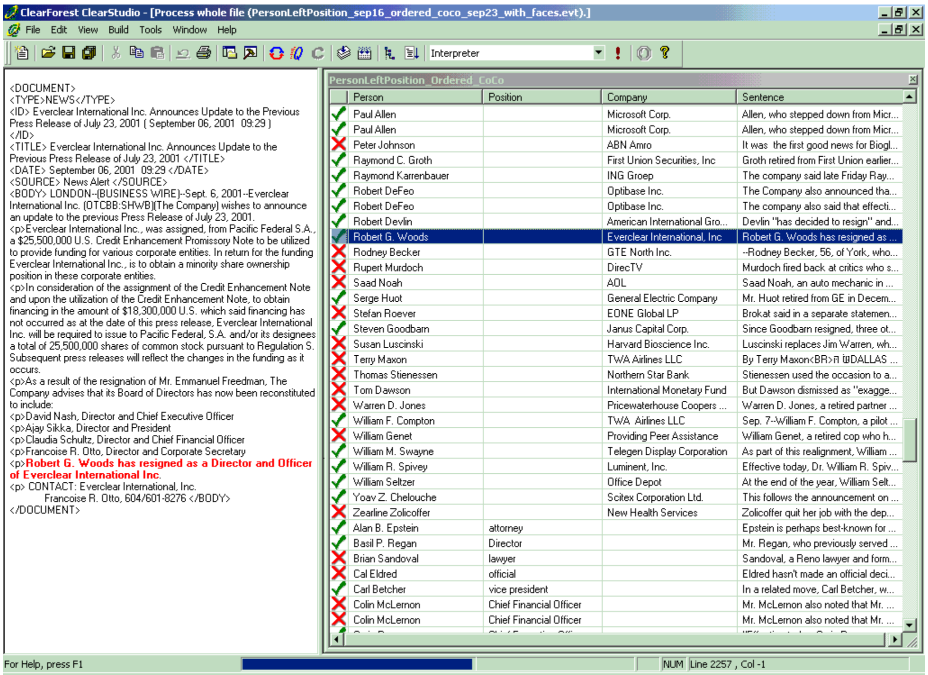
Method	Co-occurrence	Ordered Co-occurrence	Structure-Driven
Correct Instances	244	246	135
Incorrect Instances	201	132	16
Total Instances	445	396	151
Precision	54.8%	62.1%	89.4%
Recall	93.8%	94.6%	51.9%

Table 3. Acquisition template results (recall is relative to a total of 260 events)

Method	Co-occurrence	Ordered Co-occurrence	Structure-Driven
Correct Instances	353	266	174
Incorrect Instances	250	165	44
Total Instances	603	431	218
Precision	58.5%	61.7%	79.8%
Recall	97.2%	73.2%	47.9%

Table 4. Person-Left-Position template results (recall is relative to a total of 363 events)

Remark: The recall rates given in tables 3 and 4 are relative to the total number of correct instances found during the assessment of the extraction results. Recall of naïve co-occurrence does not reach 100% because this method sometimes picks up the wrong pair of companies, missing out the correct pair in the same sentence. The alternative to picking up only a single pair would be to extract *all* pairs, but that would clearly result in much poorer precision rates, as typically *½n(n-1)-1* incorrect instances would be automatically extracted from each sentence with an Acquisition keyword containing *n* companies.


ClearForest ClearStudio - [Process whole file (PersonLeftPosition_sep16_ordered_coco_sep23_with_faces.evt).]
File
Edit
View
Build
Tools
Window
Help
Interpreter
PersonLeftPosition_Ordered_CoCo
For Help, press F1
NUM
Line 2257 , Col -1

```
<DOCUMENT>
<TYPE>NEWS</TYPE>
<ID> Everclear International Inc. Announces Update to the Previous
Press Release of July 23, 2001 ( September 06, 2001  09:29 )
</ID>
<TITLE> Everclear International Inc. Announces Update to the
Previous Press Release of July 23, 2001 </TITLE>
<DATE> September 06, 2001  09:29 </DATE>
<SOURCE> News Alert </SOURCE>
<BODY> LONDON--(BUSINESS WIRE)--Sept. 6, 2001--Everclear
International Inc. (OTCBB:SHWB)(The Company) wishes to announce
an update to the previous Press Release of July 23, 2001.
<p>Everclear International Inc., was assigned, from Pacific Federal S.A.,
a $25,500,000 U.S. Credit Enhancement Promissory Note to be utilized
to provide funding for various corporate entities. In return for the funding
Everclear International Inc., is to obtain a minority share ownership
position in these corporate entities.
<p>In consideration of the assignment of the Credit Enhancement Note
and upon the utilization of the Credit Enhancement Note, to obtain
financing in the amount of $18,300,000 U.S. which said financing has
not occurred as at the date of this press release, Everclear International
Inc. will be required to issue to Pacific Federal, S.A. and/or its designees
a total of 25,500,000 shares of common stock pursuant to Regulation S.
Subsequent press releases will reflect the changes in the funding as it
occurs.
<p>As a result of the resignation of Mr. Emmanuel Freedman, The
Company advises that its Board of Directors has now been reconstituted
to include:
<p>David Nash, Director and Chief Executive Officer
<p>Ajay Sikka, Director and President
<p>Claudia Schultz, Director and Chief Financial Officer
<p>Francoise R. Otto, Director and Corporate Secretary
<p>Robert G. Woods has resigned as a Director and Officer
of Everclear International Inc.
<p> CONTACT: Everclear International, Inc.
          Francoise R. Otto, 604/601-8276 </BODY>
</DOCUMENT>
```

	Person	Position	Company	Sentence
✓	Paul Allen		Microsoft Corp.	Allen, who stepped down from Micr...
✓	Paul Allen		Microsoft Corp.	Allen, who stepped down from Micr...
✗	Peter Johnson		ABN Amro	It was the first good news for Biogl...
✓	Raymond C. Groth		First Union Securities, Inc	Groth retired from First Union earlier...
✓	Raymond Karrenbauer		ING Groep	The company said late Friday Ray...
✓	Robert DeFeo		Optibase Inc.	The Company also announced tha...
✓	Robert DeFeo		Optibase Inc.	The company also said that effecti...
✓	Robert Devlin		American International Gro...	Devlin "has decided to resign" and...
✓	Robert G. Woods		Everclear International, Inc	Robert G. Woods has resigned as ...
✗	Rodney Becker		GTE North Inc.	--Rodney Becker, 56, of York, who...
✗	Rupert Murdoch		DirecTV	Murdoch fired back at critics who s...
✗	Saad Noah		AOL	Saad Noah, an auto mechanic in ...
✓	Serge Huot		General Electric Company	Mr. Huot retired from GE in Decem...
✗	Stefan Roever		EONE Global LP	Brokat said in a separate statemen...
✓	Steven Goodbarn		Janus Capital Corp.	Since Goodbarn resigned, three ot...
✗	Susan Luscinski		Harvard Bioscience Inc.	Luscinski replaces Jim Warren, wh...
✗	Terry Maxon		TWA Airlines LLC	By Terry Maxon<BR>ก ฃDALLAS ...
✗	Thomas Stienessen		Northern Star Bank	Stienessen used the occasion to a...
✗	Tom Dawson		International Monetary Fund	But Dawson dismissed as "exagge...
✗	Warren D. Jones		Pricewaterhouse Coopers ...	Warren D. Jones, a retired partner ...
✓	William F. Compton		TWA Airlines LLC	Sep. 7--William F. Compton, a pilot ...
✗	William Genet		Providing Peer Assistance	William Genet, a retired cop who h...
✓	William M. Swayne		Telegen Display Corporation	As part of this realignment, William ...
✓	William R. Spivey		Luminent, Inc.	Effective today, Dr. William R. Spiv...
✓	William Seltzer		Office Depot	At the end of the year, William Selt...
✓	Yoav Z. Chelouche		Scitex Corporation Ltd.	This follows the announcement on ...
✗	Zearline Zolicoffer		New Health Services	Zolicoffer quit her job with the dep...
✓	Alan B. Epstein	attorney		Epstein is perhaps best-known for ...
✓	Basil P. Regan	Director		Mr. Regan, who previously served ...
✗	Brian Sandoval	lawyer		Sandoval, a Reno lawyer and form...
✗	Cal Eldred	official		Eldred hasn't made an official deci...
✓	Carl Betcher	vice president		In a related move, Carl Betcher, w...
✗	Colin McLernon	Chief Financial Officer		Mr. McLernon also noted that Mr. ...
✗	Colin McLernon	Chief Financial Officer		Mr. McLernon also noted that Mr. ...

Fig. 1. Sample Screen from the ClearStudio Environment as used for the Experiment

5. Discussion and Conclusions

For both templates, the precision results for the structure-driven strategy were significantly better than the two other simpler methods. For both methods, the ordered co-occurrence strategy performed only slightly better than the naïve co-occurrence. Precision results, under the structure-driven method were better for the *Acquisition* template than for the *Person-Left-Position* template.

The structure-driven method performs always better since it filters many noisy instances, in which the searched keyword may have a totally different meaning. For example, the *Person-Left-Position* co-occurrence rules produced as an instance the following sentence ("*quit*" was one of the keywords*):*

David Oxlade, chief executive of Xenova Group Plc, the company behind the project, said the vaccine could eventually have an important role to play in helping smokers ***quit.***

Similarly, the Acquisition co-occurrence rules produced the sentence:

ZAMBA's clients have included Aether Systems, Best ***Buy****, CompuCom, GE Medical Systems, BellSouth, Hertz, General Mills, Symbol Technologies and Towers Perrin.*

The above sentence was extracted because of the "*buy*" keyword, although, clearly none of the companies mentioned have an Acquisition relationship between them.

Rather than being an isolated case, the problem exhibited by the last example of Acquisition is very common in the domain of business news, as Acquisition keywords and particularly "acquisition" is a part of the name or description of many (acquisition) companies.

One of the main recall problems of the structure-driven method occurs in sentences in which some of the entities can only be anaphorically resolved, since it is not explicitly mentioned within the pattern, as exemplified by the sentence below ("acquisition of Pifco Holding"). The naïve co-occurrence method is not sensitive to sentence structure, so it can identify the company as long as it appears somewhere in the sentence. For a structure-driven method to overcome this problem, it has to employ an anaphora resolution mechanism. We actually do employ such a mechanism but we do not implement it, for considerations of precision, for cases like the example below, where there is no explicit anaphoric expression for which an antecedent has to be sought.

Salton, Inc. (NYSE: SFP), today reported its fiscal 2001 fourth quarter and year-end results for the period ended June 30, 2001, which includes operating results from June 1, 2001 through June 30, 2001 resulting from the previously announced acquisition of Pifco Holding PLC.

We believe that the lower precision rate for the *Person-Left-Position* template is due to the fact that this concept is more complex and can be phrased in many ways, and as a result requires more rules (patterns). Sentences that discuss a retirement of a person mention several persons and / or companies, making it more difficult to extract the correct ones. Specifically, we observed that while the rules beginning with the Company (such as in the example in Section 1 above) have very good precision, the rules beginning with the Person have worse recall.

Rather surprisingly, the ordered co-occurrence strategy proved little better than the naive co-occurrence. We believe that it is mainly due the fact that while the ordered co-occurrence may rule out sentences in which the relevant keyword has a different *syntactical* function, it fails to handle more complex *semantic* patterns. A common problem we encountered is patterns involving a more complex relationship between more than two entities. For example:

John F. Hoffner, 54, replaces Charles W. Duddles, 61, who announced in March that he would retire from Jack in the Box this year.

Regarding the Acquisition event, ordered co-occurrence only improved results in case the pattern was based on a subject-verb-object pattern. For nominal patterns no such ordering is relevant[1].

[1] while "Microsoft's acquisition of Visio" has the order subject-predicate-object (like in the case of verbs), in "the acquisition by Microsoft of Visio", for instance, both arguments follow the predicate.

Both the co-occurrence and the ordered co-occurrence strategies fail to find that the first person (*Mr. Hoffner*) is *not* the retiring one, but rather *succeeds* Mr *Duddles*. The ordered co-occurrence strategy checks only that the verb group (*would retire)* follows the entity. But it cannot observe that this verb actually refers to the second person.

Clearly, the structure-driven method requires much more extensive initial work. However, the results of this paper indicate that this is necessary if one wishes to obtain high accuracy. In any given case, the cost-benefit tradeoff must be weighed, in order to decide on the best strategy for the given application.

The above results indicate that if we are interested in all the references to an Acquisition or *Person-Left-Position* template, then the co-occurrence strategy is better recall-wise, since it extracts significantly more instances. However, our analysis shows that many of those instances occur within clauses and refer to information already known from other parts of the same document or from other documents. Note that many anaphoric cases, in this domain, fall under this category as well.

To illustrate this, consider the following sentence, which may be extracted only using the co-occurrence strategy:

"My advice is to listen to Jim Kelly," he tells a colleague, referring to their boss, the retiring UPS chairman who announced Thursday he will step down at the end of the year.

The retirement of Jim Kelly was discussed several times in the financial news in August 2001, and at least once it was within a clear template that was extracted by our structure-driven rules:

"UPS Chairman and Chief Executive Jim Kelly said Thursday he will retire".

Besides the fact that the structure-driven strategy is clearly superior over the co-occurrence strategy in precision, it also achieves a high recall rate for recent events and thus is preferable for practical applications that aim to extract precise information from news.

REFERENCES

1. Appelt D. E., Hobbs J., Bear J., Israel D. and Tyson M., 1993. "FASTUS: A Finite-State Processor for Information Extraction from Real-World Text", Proceedings. *IJCAI-93*, Chambery, France, August 1993.
2. Aseltine J., 1999. "WAVE: An Incremental Algorithm for Information Extraction". *In Proceedings of the AAAI 1999 Workshop on Machine Learning for Information Extraction.*
3. Feldman R., Liberzon Y, Rosenfeld B., Schler J. and Stoppi J., 2000. "A Framework for Specifying Explicit Bias for Revision of Approximate Information Extraction Rules". *KDD 2000*: 189-199.
4. Lin D. 1995. University of Manitoba: Description of the PIE System as Used for MUC-6 . In *Proceedings of the Sixth Conference on Message Understanding (MUC-6)*, Columbia, Maryland.
5. Soderland S., 1996. "Learning Text Analysis Rules for Domain-specific Natural Language Processing". *Ph.D. thesis*, *technical report UM-CS-1996-087* University of Massachusetts, Amherst.
6. Soderland S., Fisher D., and Lehnert W., 1997. "Automatically Learned vs. Hand-crafted Text Analysis Rules". *CIIR Technical Report.*