Entity Profile Extraction from Large Corpora*

Wei Li, Rohini Srihari, Cheng Niu, Xiaoge Li

Cymfony Inc., 600 Essjay Road, Williamsville, NY 14221, USA
{wei, rohini, cniu, xli}@Cymfony.com

Information Extraction (IE) has two anchor points: (i) entity-centric information leads to an Entity Profile (EP); (ii) action-centric information leads to an Event Scenario. Based on a pipelined architecture which involves both document-level IE and corpus-level IE, a multi-level modular approach to EP extraction from large corpora is described: (i) named entity tagging; (ii) three-level pattern matching for extracting the underlying correlated entity relationships; (iii) co-reference; (iv) document-internal merging of entity relationships into discourse EPs; and (v) cross-document fusion of EPs. The approach achieves around 90% precision and 50%-70% recall for major EP relationships. The significance of EP enhanced by cross-document fusion is demonstrated.

Key words: Information Extraction, Relationship Extraction, Information Fusion, Named Entity, Entity Profile

1 INTRODUCTION

The last decade has seen great advance and interest in the area of Information Extraction (IE). In the US, the DARPA sponsored Tipster Text Program (Grishman 1997), the Message Understanding Conferences (MUC) (Chinchor 1998), DARPA's Evidence Extraction and Link Discovery (EELD http://www.darpa.mil/ipto/Solicitations/CBD_01-27.html) and DARPA's Translingual Information Detection, Extraction, and Summarization (TIDES http://www.darpa.mil/iao/BAA03-23-PIP.pdf) have been driving forces for developing this technology.

MUC divides IE into distinct tasks such as Named Entity (NE), Template Element (TE), Template Relation (TR), and Scenario Templates (ST) (Chinchor & Marsh 1998). NE aims at the identification and classification of proper names (person, organization, location, etc.) as well as tagging time, date and numerical items (currency, percentage). TE centers around an entity, including information about its name and aliases, type and sub-type (e.g. person, military), and descriptors. TR centers around a relationship (e.g. EMPLOYEE_OF is a relationship linking a person TE and an organization TE). ST is designed to represent elaborate events in a specific domain. The most successful IE efforts to date utilize in NE technology. Systems such as *NetOwl* (Krupka & Hausman 1998), *IdentiFinder* (Miller *et al.* 1998) and *InfoXtract* (Srihari *et al.* 2000a) (Niu *et al.* 2003a) have reached near human performance, with about or above 90% for precision and recall. These systems have been either commercialized or deployed for various applications. On the other end, deep level extraction of an elaborate scenario of events such as ST is both too ambitious for commercial application and too domain dependent to allow for general application (Srihari *et al.* 2003) (Li & Srihari 2000a). TE/TR extraction has produced reasonable results, with state-of-the-art performance at around 80% F-score (precision and recall), close to the stage of being deployable (MUC-7 1998).

* This work was partly supported by grants from the Air Force Research Laboratory's Information Directorate (AFRL/IF), Rome, NY, under contracts F30602-02-C-0156 and F30602-01-C-0035. The authors wish to thank Carrie Pine of AFRL for supporting and reviewing this work.

The work on Entity Profile (EP) extraction is an extension of this line of research. EP extraction is proposed as a significant intermediate level IE task between NE and Event Extraction, collecting information about a given entity, say, *Julian Hill*, and generate his profile. The extracted profile in effect represents a miniature résumé of the person, as shown in the EP popup in our IE-supported intelligent text browser (Figure 1).

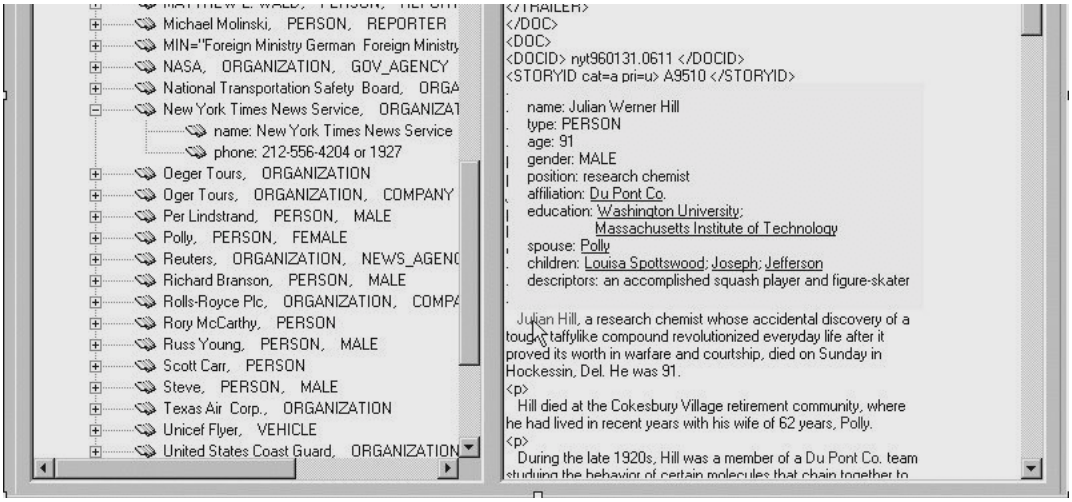

Figure 1. *Screenshot for Intelligent Browsing Prototype Based on InfoXtract EP*

2 CONCEPTUAL DESIGN

This section presents the conceptual design behind EP. The EP definition and the rationale behind the inception of EP extraction are presented first. An IE system architecture is then presented to put this research in context.

2.1 Definition of EP

Although an infinite number of relationships can occur between entities, there are certain relationships that are more predictable and relatively permanent, with respect to *Temporal Granularity* (Hobbs & Israel 1994), than others. For example, an *organization* entity is often correlated with entities like *location*, *address*, *product*, *person* (Table 1). An EP is defined as an Attribute Value Matrix (AVM) at a domain independent level for five types of entities: *organization, person, product, location, named-event* (such as historical events, conferences). The definitions can be modified and extended for a given domain to suit an IE application.

EP is an information object representing a real-world object named by a word string. Each defined relationship is represented by an attribute slot in the EP AVM. Each attribute-value pair gives some information about the entity in one aspect. The goal for EP extraction is to fill the slots for EP AVMs if such information exists in the processed text.

The progress from NE to EP is a significant development in IE representation for an entity. EP enriches the information contained in MUC TE and TR (Li & Srihari 2000). The design is to integrate the two types of information and represent them in an entity-centric format. This is implemented based on a task called Correlated Entity (CE) relationship extraction that originates from two tracks of IE tasks, namely, TE and TR in MUC. CE relationship extraction decodes sentence-internal entity-centric relationships such as *affiliation*, *position*, *age, modifiers* or *descriptors* from an NE to another unit (NE or a token string). As building blocks of EP, CE relationships consist of three types of information: (i) specific relationships such as *affiliation/staff* (corresponding to EMPLOYEE_OF in MUC TR), (ii) general relationships such as *descriptors*,

modifiers and *associated-entities*, and (iii) links to *involved-events*. The introduction of 'modifiers' and 'associated-entities' in addition to the MUC 'descriptors' reflects the desire to form a relationship *back-off* in order not to miss potential important information about an entity. For example, even if the 'head-of' relationship is not specifically defined, the 'associated-entities' relationship should still link 'bin Laden' with 'Al-Qaeda'.

Table 1. *Definition for Organization EP*

Attribute	*Appropriate value*	*Comments*
name	NE(organization)	Unless otherwise specified in a user configuration file, it is the longest NE among all alias-co-referenced NEs
aliases	NE(organization)	all co-referenced NEs except 'name'
staff	Person EP	Reverse relationship of *affiliation*
head	Person EP	Reverse relationship of *head-of*
location	Location EP	Reverse relationship of *location-of*
products	Product EP	Reverse relationship of *product-of*
revenue	NE(money)	
found-time	NE(time)	
mother-organization	Organization EP	Reverse relationship of *children-organizations*
children-organizations	Organization EP	Reverse relationship of *mother-organization*
address	NE(address)	
phone	NE(phone)	
email	NE(email)	
url	NE(url)	
descriptors	Token-string	Typically a noun phrase
modifiers	Token-string	Typically an adjectival phrase
associated-entities	EP	Pointer to EP
involved-events	Event	Pointer to Event

CE relationship extraction is implemented by the cascaded application of a list of pattern matching grammars, to be illustrated shortly. Pattern matching methods are widely used for IE tasks for their superior efficiency and convenience (Krupka & Hausman 1998; Hobbs 1993; Silberztein 1998; Li & Srihari 2000a; Aone & Ramos-Santacruz 2000). More recently, we have also explored bootstrapping techniques for relationship extraction using only a raw corpus and a few "seeds" (Niu *et al.* 2003b).

2.2 System Architecture

The InfoXtract IE system architecture in Figure 2 (Srihari *et al* 2003) distinguishes three layers : IE Engine, (cross-document) IE Fusion, and IE Applications. These three layers are linked by IE Repository. For the IE Engine, a document is the largest unit for processing; this defines the scope for document-internal information merging based on discourse analysis. The output of the engine processing goes to IE Repository. IE Fusion is designed to perform cross-document consolidation of the extracted information objects in the repository for a given archive. Finally, an IE-supported application will access IE Repository to provide information services such as Entity Tracking in a variety of IE-supported application areas such as question answering (QA) (Srihari & Li 2000b) (Li *et al.* 2002), intelligent browsing and navigation (Figure 1 and Figure 4), information visualization, automatic summarization, etc.

This three-layer design enables the engine and its extracted information objects to be exploited by various applications. The application-specific development can therefore be kept independent of the engine development, with IE Fusion mediating between the two.

At the engine layer, a hierarchical, multi-level pipeline architecture is used for extracting three major types of targeted information objects, namely, NE, EP and Event. This architecture is based on the design philosophy of strict modularization and component technology. Each lower level component can be developed independently, and runs independent of the higher level components.

For example, the deployment of EP extraction does not have to wait until the event extraction component (Pragmatic Filtering and Event Merging) is in place.

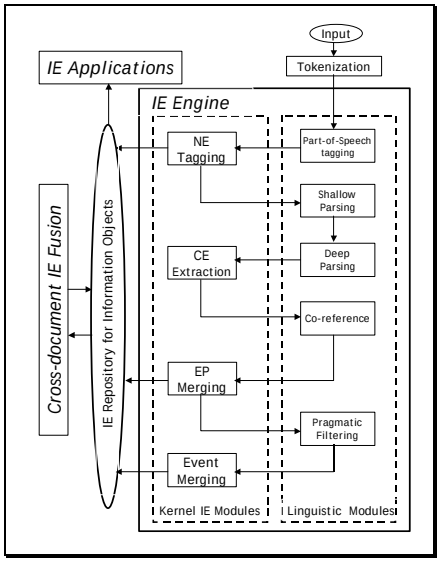

Figure 2. *Overall IE System Architecture*

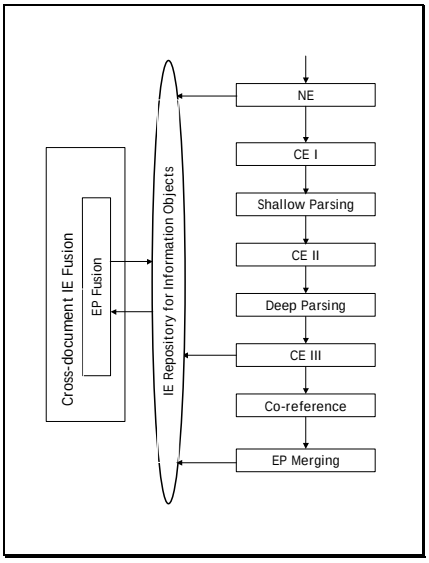

Figure 3. *EP Extraction Modules*

3 ENTITY PROFILE EXTRACTION PROCESS

3.1 Document-level IE: Relationship Extraction

This section details the process of the document level IE following the data flow shown in Figure 3. Figure 3 is a detailed view of the overall IE architecture in Figure 2, highlighting three CE modules and other supporting modules that contribute to EP.

Named Entity tagging is the starting point for EP. A high-performance NE tagger is vital for EP extraction since NEs are anchoring points for extracting CE relationships. We adopt a hybrid NE system with over 90% precision and recall (Srihari *et al.* 2000a) (Niu *et al.* 2003a) for this research.

The next IE objective is the extraction of CE relationships. It is effective and handy to develop multi-level CE grammars based on different needs for structural support from linguistic processing. More specifically, three levels of CE relationship extraction are identified.

CE I handles the most local phenomena based on pattern-matching a linear string of tokens and NEs. Two sample rules are formulated below:

NE1(ORG) + N2(position) + NE3(PER)
⇨ affiliation<NE3, NE1>;
⇨ position<NE3, N2>
NE1(LOC) + '-based' + NE2(ORG)
⇨ location<NE1, NE2>

The first rule links a person NE (NE3) with an organization NE (NE1) with the *affiliation* relationship; it also extracts a position noun (N2) like 'spokesman', 'chairman', 'secretary', 'researcher', 'salesman', etc. as the fill to the *position* attribute slot for the person NE (NE3). This rule covers cases like 'UAW spokesman Owen Bieber'. The second rule works for cases like 'Seattle-based Microsoft'.

For this type of very local phenomena, parsing is unnecessary. In fact, a shallow parser would group both 'UAW spokesman Owen Bieber' and 'Seattle-based Microsoft' as basic Noun Phrases (NPs). Therefore, the proper structural basis for CE I is identified to only require NE tagging.

Shallow Parsing matches patterns of simple, un-embedded linguistic structures and groups them into basic units. The major basic units are NPs such as 'John Smith', 'Cymfony Inc.', 'a beautiful girl', 'the most advanced software system', Verb Groups (VG) such as 'has acquired', 'to be implemented', 'would have been finished', basic Adjective Phrase (AP), and basic Prepositional Phrases (PPs) such as 'to John Smith', 'for this lady'. Our work on shallow parsing is similar to Hobbs (1993) and Appelt *et al* (1995) who applied a local grammar for NP and VG shallow parsing in order to support IE tasks.

CE II is a module based on shallow parsing. The following is a pattern rule at this level for the relationships *affiliation* and *position*.

NE1(PER) + ',' + NP2(position) + PP('of'/'for'/'with'/'in'/'at', NE3(ORG))
⇨ affiliation<NE1, NE3>;
⇨ position<NE1, NP2>

This rule applies to cases like 'Robert Callahan, spokesman of Seattle-based Microsoft'. Note that the relationship *affiliation* between 'Robert Callahan' and 'Microsoft' in the preceding example cannot be captured without the structural basis provided by Shallow Parsing. This is because the last organization NE (NE3) has a preceding modifier 'Seattle-based';[1] pattern matching has to jump over (ignore) such modifiers in order to find the related entity. This 'jump-over' operation is difficult to realize when shallow parsing is not available. This is because pre-modifiers can also take various forms of various lengths.

Deep Parsing decodes logical subject-verb-object and other grammatical relationships such as *modification*, *conjunction* and *apposition*. Deep parsing in our system has two critical features: (i) unlike conventional full parsing which is typically based on powerful, but often computationally costly grammar formalisms like Context Free Grammar (CFG), our parser is implemented by employing the cascaded application of a series of finite state automata on top of shallow parsing constructions; (ii) it parses the shallow parsing results of a sentence directly into its logical form represented by binary dependency links between linguistic units. These links include Verb-Subject (V-S), Verb-Object (V-O), Verb-Complement (V-C), Head-Modifier (H-M), Conjunctive Structures and Appositive Structures. This provides a common structural basis for supporting high-level IE for events as well as CE relationships.

CE III is designed to handle local CE phenomena at the clause level using keyword-driven, structure-based pattern matching rules. 'Structure' here refers to the dependency links of logical grammatical relationships provided by the deep parser. The underlying formalism for structure-based grammars has been extended from the conventional finite state formalism in that pattern matching happens at the structural level instead of at the linear string level. A sample rule driven by the keyword 'appoint' is given below.

'appoint': V-S: NP1(ORG)
V-O: NP2(PER)
V-C: PP('as', N3(position))
⇨ affiliation<NP2, NP1>;
⇨ position<NP2, N3>

This simple and intuitive rule covers phenomena in the canonical form 'some organization appoints some person as some role/title'. Since structural variations such as passive patterns are already consumed by the deep parser, this simple rule is powerful enough to cover all the following cases:[2]

[1] Our Finite State Automata Toolkit provides the functionality of matching a PP by checking the constraints on both the preposition node ('of'/'for'/'with'/'in'/'at' in this case) and the semantic head node (NE3 in this case).

[2] Although at any given moment, it is not guaranteed that the deep parser will always be able to decode all the structural variations into *logical form*, however, the level III CE performance is automatically enhanced with

IBM just appointed John Smith as its CEO.
Abc Inc., which has been reporting losses for three consecutive quarters, has appointed Peter Lee as its new CEO.
He was recently appointed by Xyz Corp. as its CEO.
Recently appointed as CEO by this Internet start-up, he was expected to ……
……

Co-reference consists of two sub-modules: (i) alias co-reference: e.g. 'Bill Clinton' with 'William J. Clinton', 'IBM' with 'International Business Machine'; (ii) anaphoric co-reference: decoding the NE referents for pronouns and anaphors, such as 'he' for 'Bill Clinton' and 'this company' for 'IBM'. Alias Co-reference is much more tractable than resolving anaphors. Manual checking shows that the Alias Co-reference module performs with 90-95% accuracy for person names and organization names. As for anaphoric co-reference, a hybrid model is being developed and benchmarked.

EP Merging is designed to merge multiple locally extracted CE relationships involving a given entity into its discourse EP. This is accomplished with support from Co-reference. This module is also responsible for linking discourse EPs and their related events together. The results are richer and more condensed, as shown in the sample EP in Figure 1 previously. There is a built-in option for using the entire CO support or only the more reliable Alias Co-reference support. This helps to balance IE precision and recall to suit different application needs. For the experiments reported in this paper, we have only used the more reliable alias co-reference support. EP Merging ends the engine layer processing of EP extraction. The results are discourse EPs output to IE Repository.

3.2 Corpus-level IE: EP Fusion

It is a significant step for IE to proceed from the document level to the corpus level. The InfoXtract IE Repository system can handle multi-gigabyte corpora and its processing results in support of corpus-level IE and text mining. A proprietary indexing scheme has been developed that enables fast querying over both the linguistic structures and keyword strings as well as statistical similarity queries.

During the course of the EP research, we found that the CE relationship information is by nature 'sparse data' (Table 5). It is often the case that the majority of names in a document (except for biography type of documents such as the report for 'Julian Hill' shown in Figure 1) are mentioned with little further information about the entities to which these names refer. Only a few names are coupled with correlated information such as *affiliation*, *position*, *descriptors*, *modifiers*, *associated-entities*, etc. In such a situation, cross-document fusion of discourse EPs into corpus-level EPs is the key to enrich the information object to a useful level of content.

EP Fusion is a corpus-level module designed to further merge, consolidate and link EPs across documents in the repository. Merging enriches information, consolidation involves eliminating redundant information and linking connects this EP with other related EPs or events. In terms of EP extraction, this level is crucial to make the extracted information useful for supporting IE applications.

A central problem in EP Fusion is to determine if the same name in different articles refers to the same person. Following (Bagga & Baldwin 1998), we implemented a snippet clustering algorithm using vector space model. The major problem with this approach is that it cannot determine whether the distinguished clusters correspond to two individuals or refer to two distinguished contexts involving the same person. There are cases where this approach can successfully distinguish two persons using the same name, for example, one person is a sportsman and is frequently mentioned in the sports related events and the other is a politician. But more

---

each updated version of the parser. Note also that NP(ORG) etc. is implemented by a macro which checks either proper name NP (with the tag *NeOrg*) or anaphoric NP of ORG type (e.g. 'this Internet start-up').

frequently, context clustering uncovers multiple aspects of the same person. One experiment discovered two *Bill Clinton's*, in fact, two personalities of the same person: *Bill Clinton* as statesman and *Bill Clinton* as involved in the sex scandal. This shows the general weakness of clustering in terms of entity identification, that is, the cluster contours are entirely data driven. As such, the existing approach needs significant enhancement before it is sufficiently reliable to support EP fusion.

In practice, however, some simple heuristics are fairly effective. We currently use the following heuristics for cross-document EP merging: (i) for company and product EPs, merging is based on string matching of the discourse EP's attribute *name* or *aliases* because their names (or brands) are often trademarked or uniquely registered; and (ii) EPs with multi-token names/aliases are assumed to refer to the same entity as long as their NE types match, e.g. [Mr. Bill Clinton]/NeMan and [Mrs. Clinton]/NeWoman will not be merged (exceptions include extreme cases such as two *George Bush*'s in the same political domain). In addition, a user-configurable repository-level alias list for Very Important People (VIP) or other entities is checked to merge aliases that are believed to be safe for the given corpus or domain: this provides the ability for customization to a particular domain or source.

The most challenging case involves the person EPs with single-token names (this *John* is not that *John*). Fortunately, with our modular approach which performs discourse EP merging before cross-document EP fusion, this problem is not serious in practice. In most documents, news articles in particular, full names for individuals, e.g. *John Smith*, are at least mentioned once, usually in the beginning of an article when people are introduced, before single-token aliases, say, *John* or *Mr. Smith*, are used. The alias co-reference module in IE Engine can leverage this type of discourse constraint, which conforms to the general *one sense per discourse* principle (Gale *et al* 1992). The remaining single-token names/aliases that are not important enough to be included in the repository-level alias list are currently not merged due to the risks involved. But this is minority phenomenon.

To summarize, the corpus-level IE has two anchor points: (i) entity-centric information which leads to an EP; (ii) action-centric information which leads to an Event Scenario (ES). Our experiments show that EP Fusion is a more tangible task than *ES Fusion* thanks to the availability of effective heuristics and the alias support.

4 EXPERIMENTAL RESULTS AND BENCHMARKING

We have conducted experiments on two news corpora regarding international terrorism. The first corpus (26.7 MB, 5,504 news articles) is drawn from the Foreign Broadcast Information Service (FBIS) sources, with the date range Jan. 2000 through Jan. 2002. The second corpus (106 MB, 23,554 news articles) is drawn from North American (NA) news sources in 2002 (NY Times, CNN, etc.).

The text from the FBIS corpus is entirely uppercase, so we used the optional case restoration module (Niu *et al.* 2003c) before IE processing. Since the ultimate corpus-level EP AVM is a fused information object, there is no easy way of directly benchmarking it using traditional annotated corpus methods. Random human checking from our implemented web-based prototype on the two corpora show that most of the extracted EP information makes sense to the associated entities despite a degree of noise (due to the lack of case information, the FBIS EP results contain more noise than the NA results).

The quality of EP is directly affected by the underlying support from Case Restoration (for FBIS), NE, Alias Co-reference and Parsing as well as CE Extraction. We have run a series of benchmarking tests using blind testing corpora in the news domain, with the focus on person and organization entities. In addition to the high 90's performance for Case Restoration (98% precision/recall combined F-score) and Alias Association (94% F-score), the benchmarks for other major components are described below.

An annotated testing corpus of 177,000 words is used for measuring NE Performance for both normal case-sensitive input and case-insensitive input (Table 2), using an automatic scorer following MUC NE standards (P for Precision, R for Recall and F for F-score). The overall F-score for NE is around 90% for both scenarios.

Table 2. *InfoXtract NE Benchmarking*

Type	Case-sensitive			Case-insensitive		
	P	R	F	P	R	F
ORG	89.0%	87.7%	88.3%	84.4%	83.7%	84.1%
PERSON	92.3%	93.1%	92.7%	91.2%	91.5%	91.3%

From the InfoXtract-processed testing corpus drawn, we randomly pick 250 logical Subject-Verb-Object (SVO) structural links and 60 AFFILIATION and POSITION relationships for manual checking the performance of deep parsing and CE extraction (Table 3) by non-developer linguists.

Table 3. *Parsing/CE Benchmarking I*

	SVO		CE	
Case	sensitive	insensitive	sensitive	Insensitive
P	89.50%	89.86%	96.0%	93.5%
R	81.67%	81.25%	82.8%	74.1%
F	**85.41%**	**85.34%**	**88.9%**	**82.7%**

Table 4. *Parsing/CE Benchmarking II*

	P	*R*	*F*
Shallow parsing	95.3%	96.7 %	**96.0%**
Deep parsing	83.3%	79.3%	**81.3%**
CE-Affiliation	92.3%	53.3%	**72.8%**
CE-Position	91.3%	70.0%	**80.7%**
CE-Location	83.3%	50.0%	**66.7%**
CE-Descriptors	58.3%	53.8%	**56.1%**
CE-others	60.4%	48.7%	**54.6%**

In addition to our own testing corpus, we also used the MUC-7 *dry run* corpus (also in the news domain) in further benchmarking parsing and CE (Table 4) for the normal case-sensitive input. This time, we have included shallow parsing and additional deep parsing relations than SVO (Verb-Complement relation, Head-Modifier relation, Equivalence Relation and Conjunctive relation) as well as other CE relationships. The results in Table 3 and Table 4 are fairly consistent, reflecting the current status of the InfoXtract capabilities.

The CE benchmarks in Table 3 and Table 4 are close or comparable to the best systems for the corresponding TE/TR tasks in the MUC community: the best MUC performance is 75.6% F-score for TR and 86.8% F-score for TE (Chinchor 1998). The low score for 'Descriptors' is partially due to the requirement of better coordination in making a bigger descriptor phrase from basic NP and PP units. A review of the failed cases shows that the F-score can be raised to over 75% easily. The modest score for the remaining CE relationships ('others') beyond MUC is found to be associated with two inter-related factors: (i) some relationships are 'sparse' in real-life data; a corpus with sufficient instances is not yet available in guiding the rule development; (ii) some CE grammars are under-developed, some simple and clear patterns are not built into the grammars yet, e.g. the 'spouse'-grammar did not have rules driven by the keyword 'marry'. This is our first iteration of completing the development-benchmarking loop for the entire EP extraction process. There is considerable room for improvement given a few more iterations of the system development.

Generally speaking, the keyword-driven, multi-level rule system for CE is geared more to precision than recall. Our observation is that it is fairly easy to reach 85-90% precision and 50%-70% recall for initial development once a domain is determined. To achieve between 50% and 70% for recall, it is basically the size of the grammars that matters. When more time is spent in the development of more pattern rules, the recall will pick up gradually. But going beyond 70% recall is difficult. The majority of CE relationship instances are expressed with fixed or easily predictable patterns in a given domain while the remaining instances can take various forms in both vocabulary and structures. If we narrow our general news domain to some more specific domain, the recall is expected to rise.

Due to the information redundancy in a large corpus, even with a modest recall at about 50%-70%, the EP extraction system demonstrates tremendous value in collecting information about an entity, as shown in the sample EP for 'Burhanuddin Rabbani' extacted from the NA corpus in the Figure 4 screenshot of our implemented prototype for cross-document EP. The extracted EP centralizes a significant amount of valuable information about this organization.

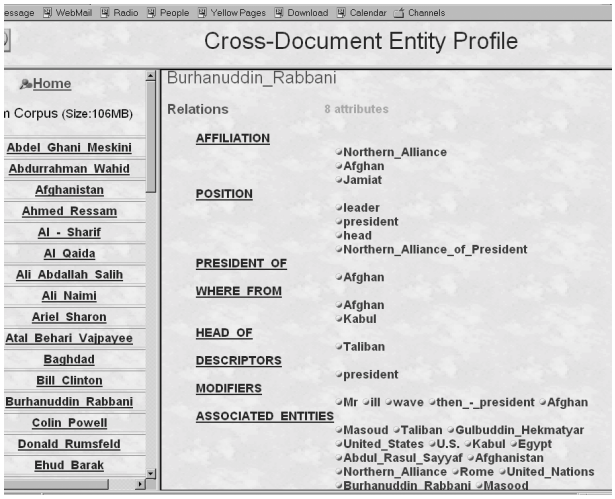

Figure 4. *Screenshot for Cross-document EP Prototype*

The richness of information for the extracted EPs on the two corpora is illustrated in Table 5, in which 'Discourse EPs' (number of document-level EPs), 'Corpus EP' (number of corpus-level EPs after fusion), 'CE' (total number of distinctive CE relationship messages), 'Events' (total number of involved events), 'Message/EP' (average number of distinct messages per corpus EP), 'Top 10' (average number of distinct messages per corpus EP for the top 10 information-rich EPs, which include *Osama bin Laden, George W. Bush, Colin Powell, Saddam Hussein, Taliban, United Nations, Pentagon* for the NA corpus and *Kim Chong-il, Vladimir Putin, United Nations, European Union, Taliban* for the FBIS corpus).[3] In a recent experiment involving processing a 1.2GB corpus containing ~100,000 news articles and a total of ~88,000,000 words, we randomly selected several big names for EP fusion and retrieval. They all resulted in huge profiles, e.g. the *Microsoft* EP contains about 2,000 distinctive messages. Without the EP extraction technology, all this potential useful information would remain hidden in huge archives. Information extraction of EPs complements the traditional Information Retrieval (IR) techniques in addressing *definition questions* such as 'Who is Kim Chong-il?' or 'What is Taliban?' (Srihari & Li 2000b) and in assisting the tracking of entities (e.g. terrorists in a watch list).

Table 5. *EP Information Statistics*

Corpus	EP Type	Discourse EPs	Corpus EPs	CE	Events	Message/EP	Top 10
FBIS (26.7MB)	Person	15,666	9,571	12,864	11,718	2.57	109
	Org.	35,468	16,827	24,858	10,758	2.12	41
NA (106MB)	Person	131,769	55,377	108,921	94,806	3.68	472
	Org.	218,310	60,922	132,301	72,066	3.35	417

[3] A distinct message is defined as a unique attribute-value pair in the cross-document fused EP AVM. Note that the fusion involves elimination of redundant messages, currently implemented by sub-string matching of the head units of phrases. Research is in progress in exploring algorithms that can identify same messages expressed in different words/patterns, in our concept-based IE project.

5 CONCLUSION

This paper establishes Entity Profile extraction as a significant IE task, which is a logical development of the MUC tradition. The significance of EP enhanced by cross-document fusion is demonstrated. A modular, multi-level approach is presented to show how this work is done with the support of various linguistic and IE modules. It leads to a high-precision EP extraction system. The modest recall for extracting the underlying correlated entity relationships is compensated during the cross-document EP fusion due to the information redundancy in a large corpus.

Future direction includes exploring efficient and systematic ways of enhancing relationship extraction recall and more sophisticated cross-document entity co-reference. In addition, we have been focusing on developing IE domain porting tools that will facilitate EP customization, based on bootstrapping (Niu *et al.* 2003a, 2003b) and Lexicon Grammar Development Environment which enables example-based semi-automatic rule writing (Srihari *et al.* 2003).

REFERENCES

Aone, A. & M. Ramos-Santacruz 2000. REES: A Large-Scale Relation and Event Extraction System. *Proceedings of ANLP-NAACL 2000*, Seattle.

Bagga, A. & B. Baldwin 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. *Proceedings of COLING-ACL'98*: 79-85, Montreal.

Chinchor, N. & E. Marsh 1998. MUC-7 Information Extraction Task Definition (version 5.1). *Proceedings of MUC-7*.

Chinchor, N. 1998. OVERVIEW OF MUC-7/MET-2. *Proceedings of MUC-7*.

Gale, W., K. Church & D. Yarowsky 1992. One Sense Per Discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop*: 233-237.

Grishman, R., 1997. TIPSTER Architecture Design Document Version 2.3. Technical report, DARPA.

Hobbs, J.R. 1993. FASTUS: A System for Extracting Information from Text. *Proceedings of the DARPA Workshop on Human Language Technology*: 133-137, Princeton, NJ.

Hobbs, J.R. & D. Israel 1994. Principles of Template Design. *Proceedings of Human Language Technology Workshop*: 177-181, NJ.

Krupka, G.R. & K. Hausman 1998. IsoQuest Inc.: Description of the NetOwl (TM) Extractor System as Used for MUC-7. *Proceedings of MUC-7*.

Li, W, R. Srihari, X. Li, M. Srikanth, X. Zhang & C. Niu. 2002. Extracting Exact Answers to Questions Based on Structural Links. *Proceedings of Multilingual Summarization and Question Answering* (COLING-2002 Workshop), Taipei, Taiwan.

Li, W & R. Srihari 2000. A Domain Independent Event Extraction Toolkit, Phase II Final Technical Report. Air Force Research Laboratory, Rome Research Site, New York.

Miller, S. *et al.* 1998. BBN: Description of the SIFT System as Used for MUC-7. *Proceedings of MUC-7*.

Mohri, M. 1997. Finite-State Transducers in Language and Speech Processing. *Computational Linguistics*, Vol.23, No.2: 269-311.

Niu, C., W. Li, J. Ding & R. Srihari 2003a. A Bootstrapping Approach to Named Entity Classification Using Successive Learners. *Proceedings of ACL-2003*. Sapporo, Japan.

Niu, C., W. Li, R. Srihari & L. Crist 2003b. Bootstrapping a Hidden Markov Model for Relationship Extraction Using Multi-level Contexts. *Proceedings of* PACLING'03. Halifax, Nova Scotia, Canada.

Niu, C., W. Li, J. Ding & R. Srihari. 2003c. Orthographic Case Restoration Using Supervised Learning Without Manual Annotation. *Proceedings of the 16th International FLAIRS Conference 2003*, Florida.

Silberztein, M. 1998. Tutorial Notes: Finite State Processing with INTEX. COLING-ACL'98, Montreal (also available at http://www.ladl.jussieu.fr).

Srihari, R., W. Li, C. Niu & T. Cornell. 2003. InfoXtract: A Customizable Intermediate Level Information Extraction Engine. *HLT-NAACL03 Workshop on The Software Engineering and Architecture of Language Technology Systems* (SEALTS). Edmonton, Canada.

Srihari, R, C. Niu & Li, W. 2000a. A Hybrid Approach for Named Entity and Sub-Type Tagging. *Proceedings of ANLP 2000*, Seattle.

Srihari, R & Li, W. 2000b. A Question Answering System Supported by Information Extraction. *Proceedings of ANLP 2000*, Seattle.