Knowledge-Based Information Extraction: A Case Study of Recognizing Emails of Nigerian Frauds

Yanbin Gao[1] and Gang Zhao[2]

[1] Advanced System Development Co, Ltd, Beijing, China
yanbin_gao@asdc.com.cn
[2] STARLab, Computer Science Department, Vrije Universiteit Brussel, Belgium
gang.zhao@vub.ac.be

Abstract. This paper describes the methodology, process and results of developing an application ontology as software specification of the semantics of forensics in the email suspicious of Nigerian frauds. Real life examples of fraud emails are analyzed for evidence and red flags to capture the underlying domain semantics with an application ontology of frauds. A model of the natural language structure in regular expressions is developed in the light of the ontology and applied to emails to extract linguistic evidences of frauds. The evaluation of the initial results shows a satisfactory recognition as an automatic fraud alert system. It also demonstrates a methodological significance: the methodical conceptual modeling and specific purpose-driven linguistic modeling are effective in encapsulating and managing their respective needs, perspectives and variability in real life linguistic processing applications.

1 Introduction

This paper presents a study of developing an ontology as a specification of application semantics for linguistic engineering and processing. The application domain is the detection of the emails suspicious of Nigerian frauds. The ontology is a knowledge model of the fraud forensics. It serves as conceptual basis for specifying linguistic rules in regular expressions and information extraction to recognize fraud evidences from texts, samples of Nigerian fraud emails.

The paper introduces the ontology development methodology (Section 2), describes its development process (Section 3) and the linguistic engineering with respect to the resultant ontology (Section 4), analyses the experiment results of evidence recognition (Section 5) and concludes with a summary (Section 6).

2 Ontology Modeling Approach

We make a methodological distinction between two viewpoints in knowledge engineering for ontology-based information extraction: application semantic model vs. linguistic model. The application semantics is described as a conceptual model of basic concepts and relationships in the problem domain. Though the intended solution of information extraction is a software artifact of natural language processing, we recognize the necessity for a problem-oriented semantic model, independent of linguistic considerations, and the solution-oriented natural language models built thereupon. Besides the engineering need for a modular approach and complexity manage-

A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 161–172, 2005.

ment, it is important to capture the recurrent essential qualities of frauds underlying highly varied and dynamic linguistic expressions, in order to encapsulate and keep up with changes and evolutions efficiently.

The ontology to create is not an upper or base or foundation ontology, neither is it a domain ontology. It is an application specific ontology of the fraud forensics for recognizing Nigerian advance fee frauds. It is a part of or extension of a topical ontology of fraud forensics. It is specific in conceptual perspectives and relevance across multiple domain ontologies. Its purpose at this point is to describe what rather than how.

2.1 Ontology Representative Framework

The DOGMA (**D**eveloping **O**ntology-**G**uided **M**ediation for **A**gent) ontology representation framework defines an ontology as a set of *lexons* and their commitments in particular applications. A lexon is defined as 5-tuples, <*Context, Term*$_1$*, Role*$_1$*, Term*$_2$*, Role*$_2$ >. It represents a fact type: a relationship type between two object types (*Term*$_1$ and *Term*$_2$ playing *Role*$_1$, and *Role*$_2$) [3] [5]. While lexons capture the underlying concepts and relationships, the commitment ground them to a particular application or task requirement [2] [9] with specific constraints and instantiations.

2.2 Ontology Engineering Methodology

AKEM (**A**pplication **K**nowledge **E**ngineering **M**ethodology) is devoted to ontology-based knowledge engineering [8] [9] [10]. Based on the DOGMA ontology representation framework, it is designed for multi-disciplinary geographically distributed team of knowledge engineering. It follows a lifecycle model similar to RUP [4] through the activities of *scoping*, *analysis*, *development* and *deployment*. It emphasizes semantic scoping and traceability of decision making in modeling with specific deliverables of particular formats.

Scoping is to identify a part of the universe of discourse for modeling and development. *Stories* are used to convey business case, scenarios and their semantic scope to the knowledge or ontology engineers. Analysis is to create a knowledge constituent model to describe how the application semantics can be decomposed and how each constituent is elaborated in the description of business logic. Development is a process of creating ontologies to capture the meta knowledge: the semantic relationship underpinning the domain knowledge. It has three main tasks: *extraction*, *abstraction* and *organization* of lexons. Deployment is to specify commitments: a set of instantiated and constrained lexons with respect to specific applications.

3 An Ontology of Fraud Forensics About Nigerian Frauds

The Nigerian fraud is a type of advance fee frauds. The perpetrator seeks to rob the victim of financial resources by luring him into paying fees for fictitious administrative or financial operations in the promise of a considerable share of a fictitious capital or fortune. The potential victim is approached nowadays through unsolicited emails. One common bait is the fortune left behind by the dead without heirs. We shall illustrate our ontology modeling with this “fortune from the dead” case (FFD). It

involves a fictitious need to transfer large sums of capital into an overseas bank account. The author of emails typically claims to be a bank official or a relative of the dead and offers up to 30 percent of capital for the victim to assist the transfer. The forms of Nigerian frauds vary greatly. The underlying scheme and pattern, however, are stable and recurrent. In view of this fact, we propose to capture the underlying conceptions in terms of ontology and handle the surface variations on the basis of the conceptual model.

3.1 Knowledge Scoping and Analysis

The process follows the main stages of the AKEM methodology of scoping and analysing the application knowledge in the application domain.

3.1.1 Scoping with Stories

The semantic scope under consideration at a given time of knowledge engineering is specified and documented by a story. It not only identifies the focus or boundary of attention, but also conveys the semantic context in which it stands [8]. It is the deliverable of the scoping with a structured presentation of information. The following figure shows parts of the story and its structure.

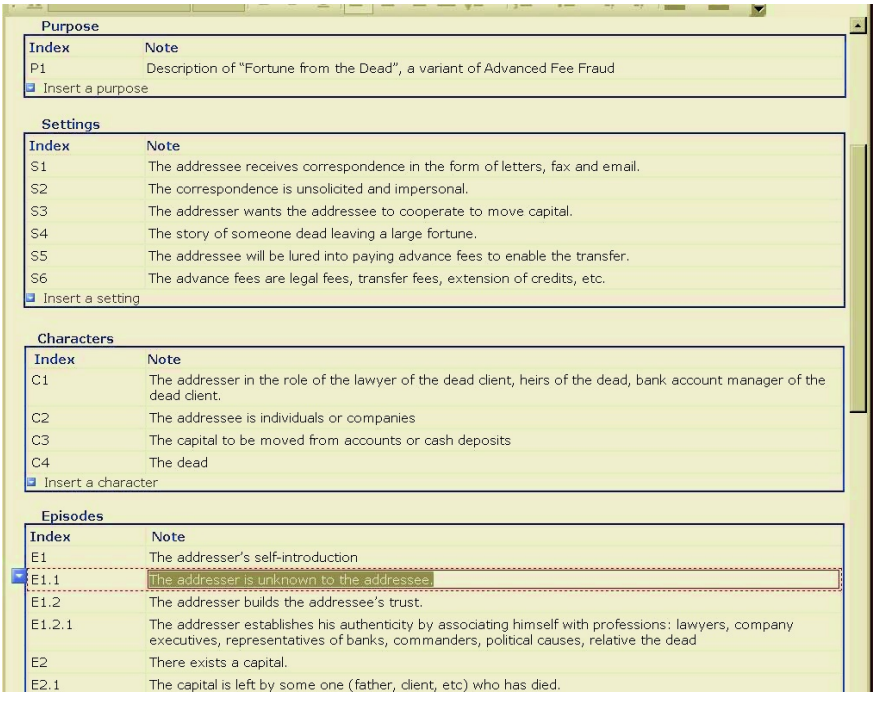

Fig. 1. Semantic scoping with the AKEM story editor.

The *Settings* of the story describe the background information whereas the *Characters* the actors or objects involved. The *Episodes* describe either sets of relationships

in hierarchy or a sequence of events or their composition. They are the starting points of the knowledge decomposition at the analysis stage of AKEM.

3.1.2 Knowledge Constituent Analysis

The analysis activity of AKEM is to create the knowledge constituent model within the semantic space defined by the story. It consists of knowledge decomposition and elaboration.

Knowledge Decomposition. It is a hierarchical structure of knowledge constituents, which is a four-layer detective model derived from Wigmore chart [7]. Fig. 2 is the knowledge decomposition of FFD. The top layer is the hypothesis of the existence of the fraud. The second layer consists of supporting postulates to the hypothesis. There are six basic ingredients of the FFD fraud scheme: medium, fraudster, victim, bait, offer and follow-up action. The third layer specifies evidences in the postulates and the fourth the facts that expresses or embodies the evidences.

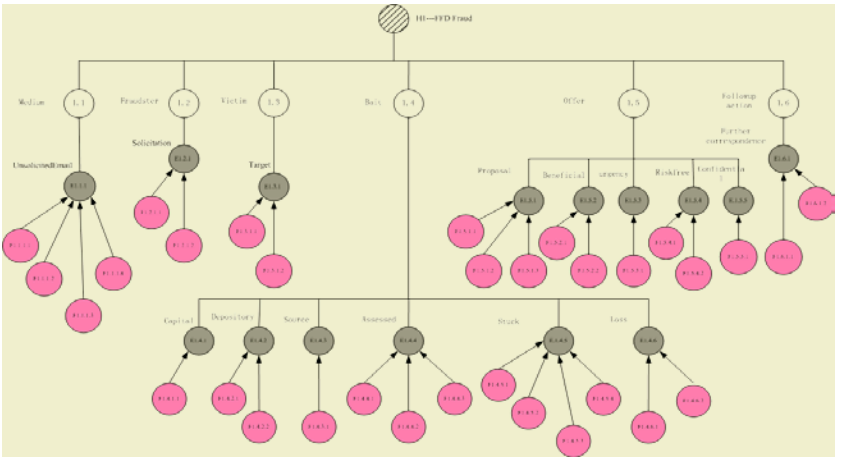

Fig. 2. Knowledge constituent model of FFD.

Knowledge Elaboration. The knowledge elaboration are the logical statements of about each constituent in a controlled language. They are a set of rules which represent the semantic connection among the knowledge constituents as well as description of the constituent. They also indicate the heuristics involved in the fraud investigation.

Fig. 3. describes the knowledge elaboration concerning Postulation Node 1.1, Evidence Node 1.1.1 and Fact Nodes 1.1.1.1, 1.1.1.2, 1.1.1.3, 1.1.1.4 and 1.1.1.5. It describes the unsolicited emails is used as medium for FFD.

3.2 Ontology Development

The ontology development seeks to define the concepts and relationships underlying the knowledge elaboration. Three tasks are performed to produce a set of lexons: extraction, abstraction and organization [10]. Extraction is to spot the key words and phrases of the elementary semantic conceptions (the left part of Fig. 4).

IF Fact (1.1.1.1) the addressee surprised by the email
OR IF Fact (1.1.1.2) the addressee is unknown addressor
OR IF Fact (1.1.1.3) the addressor introduce self
OR IF Fact (1.1.1.4) the addressor impersonal reference to addresser
OR IF Fact (1.1.1.5) the addressor describe the acquirement of addresses
THEN Evidence (1.1.1) the addressor send unsolicited email to addressee

IF Evidence (1.1.1) the addressor send unsolicited email to addressee
THEN Postulate (1.1) email is unsolicited as the medium of the fraud scheme

Fig. 3. An example of the knowledge elaboration.

Fig. 4. Extraction and abstraction of lexons in AKEM Ontology Editor.

The abstraction is a process of identifying the objects and relationships expressed the highlighted key words and phrases and formalizing them into lexons. The purpose is to model only those concepts and relationships explicitly verbalized in the deliverable of knowledge elaboration. The result of the abstraction is a set of lexons (the right part of Fig. 4). The organization is devoted to the introduction of conceptions that are presumed or implied in the story and knowledge elaboration, such as subtyping relationship. Table 1 shows some lexons produced during the ontology development phase.

Table 1. A subset of lexons concerning FFD.

Context	$Term_1$	$Role_1$	$Term_2$	$Role_2$
FFD	Fraud	CharacterisedBy	FraudType	Characterise
FFD	FraudScheme	ConsistOf	Mediums	
FFD	FraudScheme	ConsistOf	Bait	
FFD	FraudScheme	ConsistOf	Offer	
FFD	FraudScheme	ConsistOf	FollowingupAction	

4 Linguistic Modeling

The linguistic model is treated as deployment of the ontology of fraud forensics resulting from the previous activities of knowledge analysis and development. At the stage of deployment, the knowledge constituent analysis and ontology serve as development specification documents to organize and manage the linguistic engineering. The phased knowledge and language modelling encourages the formal identification of the application semantics about recurrent patterns of frauds and linguistic structures that express fraud evidences in texts, and, more importantly, the explicit representation of the mappings between two models. The layered modelling encapsulates the variation and changes in the linguistic expressions of evidences, which facilitates responsive model adaptation to keep up with the evolution of frauds.

4.1 Ontology Commitments for Recognizing FFD

The ontology model of FFD is produced without considering how and where the basic concepts and relationships are used or deployed. It focuses on the problem space rather than the solution space. The application of recognizing the "red flags" of the Nigerian frauds in emails, highlighting words and phrases expressing concepts of the fraud model.

From the deliverables from knowledge scoping and analysis, 15 red flags can be identified:

- Unsolicited email
- Targeted cooperator
- Solicitation
- The Dead
- Capital
  - Depository
  - Access
- Obstacle
  - Capital stuck
  - Capital to loose
- Proposal
  - Benefits
  - Minimal risks
  - Confidentiality
  - Urgency
- Follow-up
  - Further Correspondence

The relevant lexons are selected to create the conceptual model of the FFD. For example, the lexon, <*FFD, Addressor, Solicit, Addressee, SolicitedBy*>, shows that it is important to establish from the text or its pragmatic context that the intention of the email is to solicit. Fig. 5 is the graphic representation of some key lexons needed to produce linguistic model for red flag recognition. The conceptual model of fraud evidences in terms of relevant lexons specifies what to model in the language model.

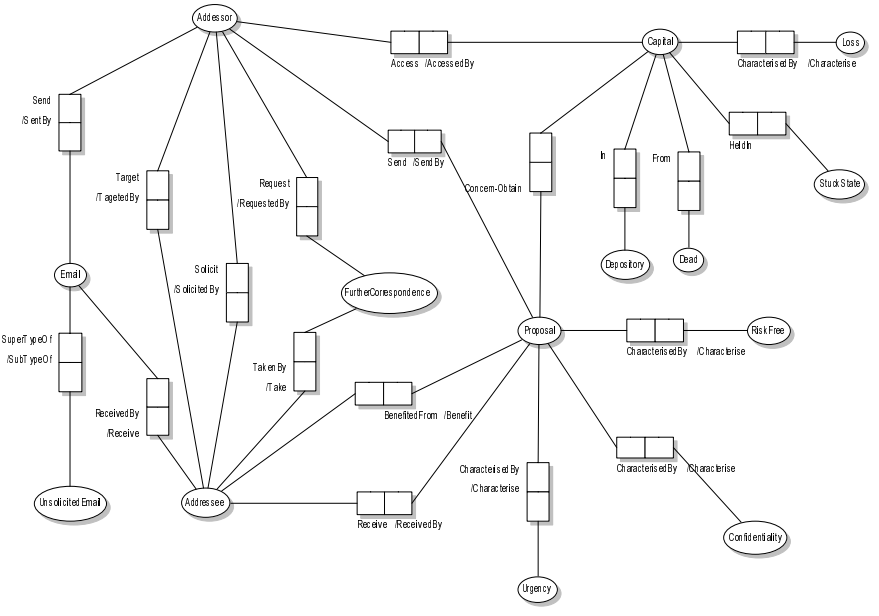

Fig. 5. Some key ontological terms and relationships in FFD.

4.2 Creating Natural Language Model

Linguistic modeling produces regular expressions of words and phrases as the linguistic evidences of frauds in emails. Since the fraud evidence is actually embodied in texts in natural language, the regular expressions are used to specify the lexical and syntactic pattern of the natural language expressions. In this study, we did not explore the use of any natural language processing/understanding engine. The popular regular expression engine was considered sufficient to explore the ontology-based natural language engineering for fraud detection at this preliminary stage. A previous similar approach of ontology-based information extraction was explored by Data Extraction Group at Brigham Young University [1].

The development process takes three steps:

- Define a semantic model based on the deliverables from the activities of knowledge analysis and ontology development)
- Describe syntactic and lexical structures that express the semantic model in the light of the representative samples
- Formalize the syntactic and lexical description in regular expressions

The linguistic model developed for the current study is based on 20 email samples of the Nigerian frauds. Table 2 shows three key intermediate results of linguistic modelling: ontological proposition to extract, linguistic tokens and the underlying linguistic structure in regular expression.

Table 2. Example deliverables in language modeling.

Lexon commitments	Natural language structures	Regular expression
"addressor seek foreigner", "addressor seek overseas firm", "addressor seek assistant", "addressor seek partnership".	need|seeking|asking|soliciting|appear|lookingfor|search|desire … foreign partner | person | assistant | help | partnership | overseas firm | cooperation |…	`(?:(?:look|seek|solicit|ask|appeal|need|desire|search|like|request)(?:ing|s){0,1}\s(?:\w*\s){0,6}(?:help|assistance|partner|participant|foreigner|person|account|permission|relationship|partnership|co{0,1}o{0,1}peration|oversea\sfirm))`

5 Detecting Fraud Red Flags

RegExTest [6] is used to extract red flags from suspicious emails against the ontology model. Fig. 6 shows how to RegExTest is used to extract the fraud evidence the suspicious email and to list and group extracted fraud evidences.

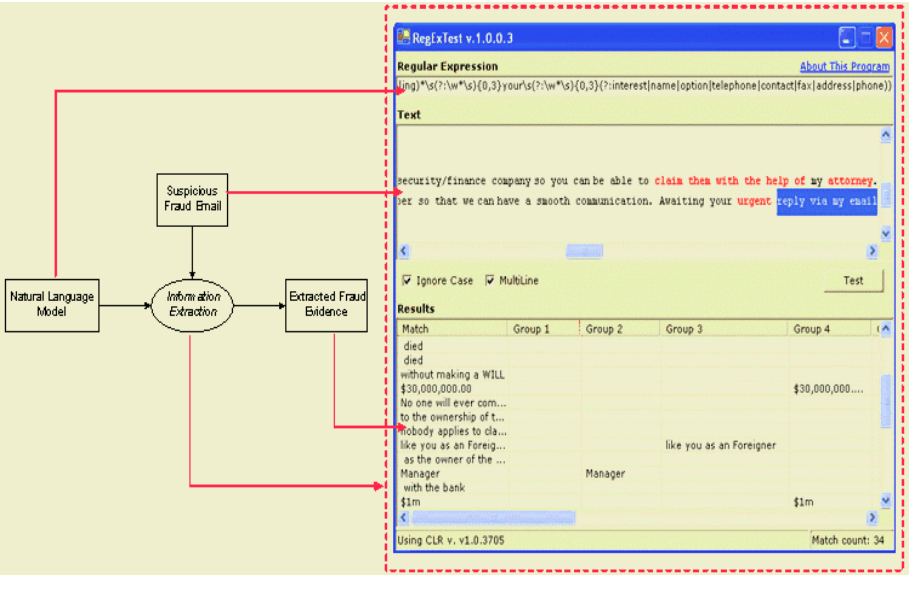

Fig. 6. Using RegExTest tool to detect fraud evidence of FFD.

5.1 Test and Evaluation

The 50 emails of FFD other than the 20 used for conceptual and linguistic modeling have been tested on the RegExTest. Table 3 shows the processing log of one email of FFD in terms of 15 *Categories of Evidence*. The *Linguistic Facts* are the linguistic tokens extracted rightly or wrongly or should have been extracted. They can be interpreted as instance one or more categories of evidence by the fraud model FFD.

Table 3. A Processing Log of An Email of FFD.

Linguistic Facts	Categories of Evidence															Total
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
Extracted	5	2	2	6	6	2	11		1	3			2			40
Mistaken	1						1			2						4
Missed			1			1									1	3

Table 4 shows a summary of the statistics of the 50 emails of FFD examined. The *Extracted* column is the total of linguistic facts extracted according to the fraud evidence and linguistic model. The *Ideal* column is the linguistic facts identified by the human investigator with respect to the fraud model. The average rate of mistaken recognition is 1.66. An average of 2.72 linguistic tokens are missed per email out of the fifty. The *Precision* counts the correct extraction in all the extraction whereas the *Recall* counts the correct recognition by the existing fraud evidence in one suspicious email. The average precision and recall of evidence recognition is 94% and 91%.

Table 4. Evaluation of 50 emails of FFD.

Email	Extracted	Correct	Mistaken	Missed	Ideal	Precision	Recall
1	40	36	4	3	39	90%	92%
2	48	47	1	4	51	98%	92%
3	44	42	2	2	44	95%	95%
4	50	49	1	1	50	98%	98%
5	53	52	1	0	52	98%	100%
6	36	35	1	3	38	97%	92%
……							
46	19	19	0	2	21	100%	90%
47	26	24	2	1	25	92%	96%
48	24	24	0	2	26	100%	92%
49	39	34	5	0	34	87%	100%
50	27	26	1	1	27	96%	96%
Average	28.54	26.88	1.66	2.72	29.6	94%	91%

5.2 Tests on Cases of "Over-Invoicing"

Though the ontology model and linguistic model is produced on FFD, we use it to extract red flags from emails suspicious of other forms of Nigerian frauds to test the model applicability and robustness. The application ontology of frauds is intended to be generic over a family of similar applications and solutions.

The case of over-invoicing is different from the FFD in that the fictitious fortune comes from over-invoicing in business contracts. 50 suspicious emails with over-invoicing cases are processed using the same ontology and language models. Table 5. summarises the statistics of the results. The average rates of mistaken and missed recognition are 1.58 and 4.32. The precision and recall of evidence recognition is 92% and 80%. The precision almost the same as the results on the FFD cases. The recall is lower, due to the lacks in the over-invoicing specific details in conceptual and linguistic model. The overall result demonstrates that the application ontology of the Nigerian fraud forensics as knowledge specification covers sufficiently the problem space and serves as a basic layer of reusable knowledge representation.

Table 5. Evaluation of 50 emails of the Over-invoicing Case.

Email	Extracted	Correct	Mistakes	Missed	Ideal	Precision	Recall
1	21	20	1	4	24	95%	83%
2	16	16	0	5	21	100%	76%
3	17	13	4	4	17	76%	76%
4	39	39	0	1	40	100%	98%
5	15	14	1	4	18	93%	78%
6	23	21	2	4	25	91%	84%
……							
45	16	14	2	4	18	88%	78%
46	11	10	1	7	17	91%	59%
47	28	25	3	2	27	89%	93%
48	27	25	2	3	28	93%	89%
49	27	21	6	2	23	78%	91%
50	24	22	2	4	26	92%	85%
Average	20.68	19.1	1.58	4.32	23.42	92%	80%

5.3 Tests on Normal Emails

The conceptual and language models are also tested on 50 non-fraudulent emails. Their subject of the emails varies from private to business or technical correspondences. The purpose of the test is to assess the differentiability between fraudulent and non-fraudulent emails by the fraud model and their corresponding linguistic model. Our assumption is that the bigger the difference is between the red flag recognition on the two sets of emails, the more performative of the model in recognizing emails suspicious of the Nigerian fraud will be.

Most linguistic tokens extracted from normal emails express the concepts of the unsolicited email, solicitation, capital and further correspondence. Expressions of the concepts relevant to the fraud model are fewer than 4 for every email in comparison with more than 10 expressions from suspicious ones.

Figure 7 shows the general trends or distribution of the three sets of tests. The percentage is the categories of evidence present out of the 15 categories in the fraud model. The linguistic tokens recognized include both the correct and mistaken recognition. The missed recognition is not considered. In other words, the distribution re-

flects the unedited evidence profile produced automatically of the suspicious and normal emails. It is however worth noting that each category of evidence carries equal weight in the percentage calculation. This is counter-intuitive, since some categories of red flags are more important than the others. Weighting on the red flags will be taken into account in the future research in order to capture experts' heuristics and intuitions.

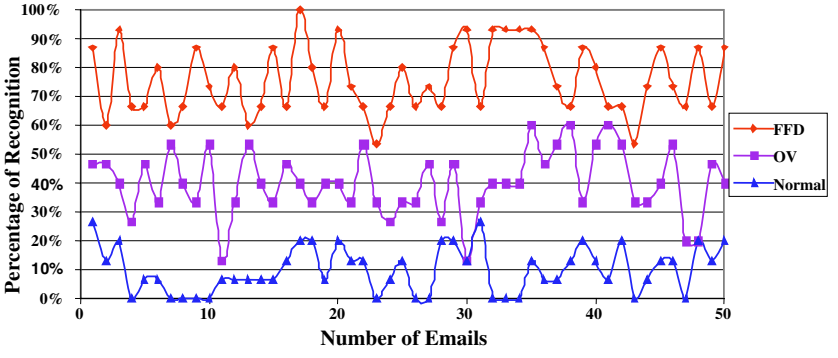

Fig. 7. Distribution of fraudulent and non-fraudulent emails.

6 Conclusion

This study focuses on the use of ontology for semantic text mining. It uses an ontology-based knowledge engineering approach to define the model of domain semantics and explores its use as software specification to guide linguistic modeling in regular expressions for information extraction. Though the solution of the automatic recognition of Nigerian frauds is linguistic processing, it is not approached as a task of linguistic engineering in the first place. Instead, the problem space is modeled as ontology-based application knowledge engineering. The decoupling of linguistic and conceptual modelling in development serves an fundamental requirement in machine assisted fraud detection. It is to capture the underlying recurrent essence of frauds while allowing for flexibility and adaptability to their creative and variable forms by distinguishing between models of problem and solution spaces and modeling knowledge in layers.

From the semantic modeling to linguistic rule specification, six major tasks are performed: problem definition, knowledge resources collection, knowledge scoping, analysis, application ontology development, and ontology deployment: natural language modeling. The whole process is purpose-driven and leads to a high precision rate.

The model of application semantics is not deployed on a natural language processing engine. Though its theoretical and potential issues are not explored, the experiment proves the pragmatic value of the dedicated and methodical domain modeling and the popular and easily available regular expression engine. A large and less structured domain can prove a serious challenge to the approach in the complexity of linguistic engineering with regular expressions. The use of easily customisable the natural language engine can be instrumental to manage the semantic and linguistic complexity.

The study also explores a methodology of development of the information extraction technology as fraud-alert expert system. The initial experiment results shows a good promise of such technology to assist monitoring and supervision of emails contents in proactive fraud detection and prevention.

Acknowledgement

The authors wish to thank Robert Meersman for his advice and support to the research on ontology. This study is partially funded by the EU 5th framework program, IST 2001-38248, the project of FF POIROT (www.ffpoirot.org).

References

1. Data Extraction Research Group at BYU. http://www.deg.byu.edu/
2. Deray, T.,Verheyden, P.: Towards a Semantic Integration of Medical Relational Databases by Using Ontologies: a Case Study. On the Move to Meaningful Internet Systems 2003: OTM 2003 Workshops, Lecture Notes in Computer Science, Vol. 2889/2003. Springer-Verlag, Heidelberg, (2003) 137-150.
3. Jarrar, M., Demey, J., Meersman, R.: On Reusing Conceptual Data Modeling for Ontology Engineering. Journal on Data Semantics, 1(1) (2003) 185 – 207
4. Kruchten, P.: The Rational Unified Process: An Introduction, Boston, Addison Wesley (2000).
5. Meersman, R.: Reusing Certain Database Design Principles, Methods and Techniques for Ontology Theory, Construction and Methodology. STARLab Technical Report 01 (2000)
6. RegExTest Project at: http://sourceforge.net/projects/regextest
7. Wigmore, J.H.: The Science of Judicial Proof as given by Logic, Psychology and General Experience. Boston, Little Brown (1937)
8. Zhao Gang: DOGMA-AKEM in FFpoirot. Draft report in WP6 of FFpoirot. (2003)
9. Zhao, G., Gao, Y., Meersman, R.: An Ontology-based Approach to Business Modeling. In: Proceedings of the International Conference of Knowledge Engineering and Decision Support (2004) 213 – 221
10. Zhao, G., Kingston, J., Kerremans, K., Coppens, F., Verlinden, R., Temmerman, R., Meersman, R.: Engineering an Ontology of Financial Securities Fraud, OTM 2004 Workshops, Lecture Notes in Computer Science, Vol. 3292. Springer-Verlag, Heidelberg, (2004) 605-620.