Automatic Ontology Creation from Text for National Intelligence Priorities Framework (NIPF)

Mithun Balakrishna, Munirathnam Srikanth
Lymba Corporation
Richardson, TX, 75080, USA
Email: {mithun,srikanth}@lymba.com

***Abstract*—Analysts are constantly overwhelmed with large amounts of unstructured data. This holds especially true for intelligence analysts with the task of extracting useful information from large data sources. To alleviate this problem, domain-specific and general-purpose ontologies/knowledge-bases have been proposed to help automate methods for organizing data and provide access to useful information. However, problems in ontology creation and maintenance have resulted in expensive procedures for expanding/maintaining the ontology library available to support the growing and evolving needs of the Intelligence Community (IC). In this paper, we will present a semi-automatic development of an ontology library for the National Intelligence Priorities Framework (NIPF) topics. We use Jaguar-KAT, a state-of-the-art tool for knowledge acquisition and domain understanding, with minimized manual intervention to create NIPF ontologies loaded with rich semantic content. We also present evaluation results for the NIPF ontologies created using our methodology.**

***Index Terms*—ontology generation, National Intelligence Priorities Framework (NIPF).**

I. Introduction

Analysts are constantly plagued and overwhelmed by large amounts of unstructured, semi-structured data required for extracting useful information [1]. Over the past decade, ontologies and knowledge bases have gained popularity for their high potential benefits in a number of applications including data/knowledge organization and search applications [2]. The data processing burden on the intelligence analysts have been relieved with the integration of ontologies to help automate methods for organizing data and provide access to useful information [3].

Though a number of applications can and have benefited due to their integration with domain-specific and general-purpose ontologies/knowledge-bases, it is very well known that ontology creation (popularly referred to as the *knowledge acquisition bottleneck* [2]) is an expensive process [4], [5]. The modeling of ontologies for non-trivial domains/topics is difficult and time/resource consuming. The *knowledge acquisition bottleneck* problems in ontology creation and maintenance have resulted in expensive procedures for maintaining and expanding the ontology library available to support the growing and evolving needs of the Intelligence Community (IC).

In this paper, we present a semi-automatic development of an ontology library for the 33 topics defined in the National Intelligence Priorities Framework (NIPF). NIPF is the *Director of National Intelligence's (DNI's) guidance to the Intelligence Community on the national intelligence priorities approved by the President of the United States of America* [6].

Lymba's Jaguar-KAT [3], [7] is a state-of-the-art tool for knowledge acquisition and domain understanding. We use Jaguar to create rich NIPF ontologies by extracting deep semantic content from NIPF topic specific document collections while keeping the manual intervention to a minimum. In this paper, we discuss the technical contributions of automatic concept and semantic relation extraction, automatic ontology construction, and the metrics to evaluate ontology quality.

II. Automatic Ontology Generation

Jaguar automatically builds domain-specific ontologies from text. The text input to Jaguar can come from a variety of document sources, including Text, MS Word, PDF and HTML web pages, etc. The ontology/knowledge-base created by Jaguar includes the following constituents:

- Ontological Concepts: basic building blocks of an ontology
- Hierarchy: structure imposed on certain ontological concepts via transitive relations that generally hold to be universally true (e.g. ISA, Part-Whole, Locative, etc)
- Contextual Knowledge Base: semantic contexts that encapsulate knowledge of events via semantic relations
- Axioms on Demand: assertions about concepts of interest generated from the available knowledge; this is useful for reasoning on text

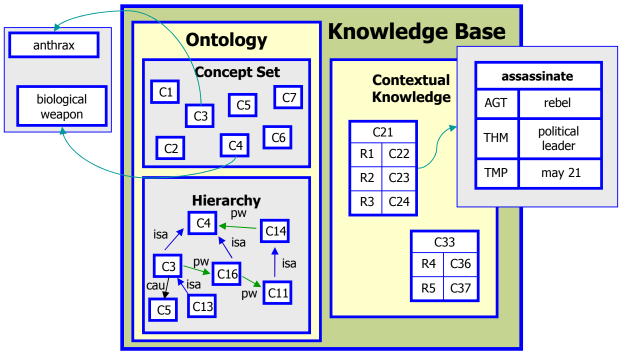
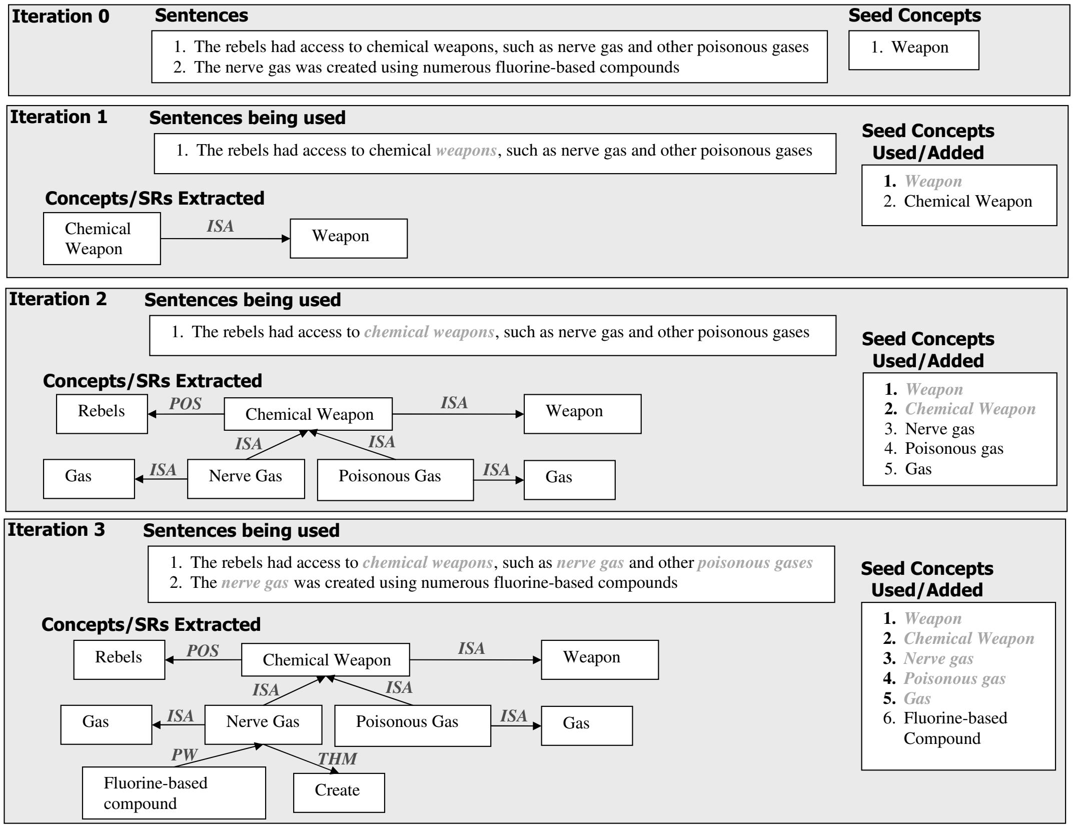


Fig. 1. An example Jaguar knowledge-base containing concepts, hierarchy and contextual knowledge.

Figure 1 shows an example Jaguar knowledge-base containing concepts, hierarchy and contextual knowledge. The



Iteration 0 — Sentences:

1. The rebels had access to chemical weapons, such as nerve gas and other poisonous gases
2. The nerve gas was created using numerous fluorine-based compounds

Iteration 1 — Sentences being used:

1. The rebels had access to chemical *weapons*, such as nerve gas and other poisonous gases

Iteration 2 — Sentences being used:

1. The rebels had access to *chemical weapons*, such as nerve gas and other poisonous gases

Iteration 3 — Sentences being used:

1. The rebels had access to *chemical weapons*, such as *nerve gas* and other *poisonous gases*
2. The *nerve gas* was created using numerous fluorine-based compounds

Fig. 2. An example depicting Jaguar's iterative process of extracting concepts and semantic relations of interest using seed concepts.

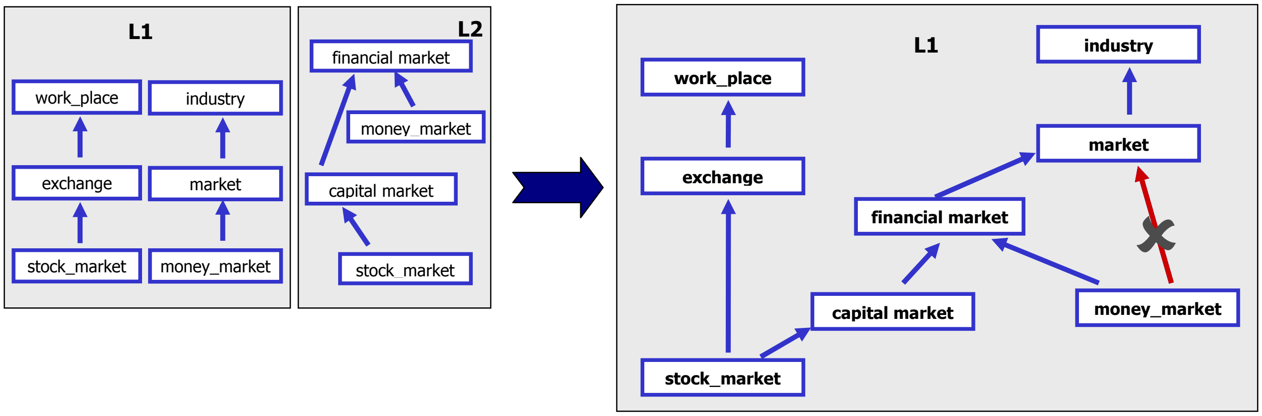


Fig. 3. An example depicting Jaguar's merging of two ontologies through conflict resolution algorithms.

input to Jaguar includes a document collection (Text, MS Word, PDF and HTML web pages, etc.) and a seeds file containing the concepts/keywords of interest in the domain. Jaguar's ontology creation involves complex text processing using advanced Natural Language Processing (NLP) tools, and an advanced knowledge classification/management algorithm. A single run of Jaguar can be divided into the following two major phases:

- Text Processing
- Classification/Hierarchy Formation

In Text Processing, the first step is to extract textual content from the input document collection. The text files then go through a set of NLP processing tools: named-entity recognition, part-of-speech tagging, syntactic parsing, word-sense disambiguation, coreference resolution, and semantic parsing (or semantic relation discovery) [8], [9]. The concept discovery module then extracts the concepts of interest using the input seeds set as a starting point and growing it based on the extracted NLP information [3].

The classification module forms a hierarchical structure within the set of identified domain concepts via transitive relations that generally hold to be universally true (e.g. ISA, Part-Whole, Locative, etc). Jaguar uses well-formed procedures [7] to impose a hierarchical structure on the discovered concepts set using the semantic relations discovered by Polaris [1] and with WordNet [10] as the upper ontology.

A. Automatically Building NIPF Ontologies

In this paper, we use Jaguar to create an ontology library for the 33 topics defined in NIPF. For each NIPF topic, we collected 500 documents from the web (the $Weapons$ topic was an exception and its collection had only 50 Wikipedia documents) and manually verified their relevance to the corresponding topic. We then use Jaguar to create an ontology, for each identified NIPF topic. Jaguar builds each ontology with rich semantic content extracted from the corresponding NIPF topic document collection while keeping the manual intervention to a minimum. These ontologies are fine-tuned to contain the level of detail desired by an analyst.

1) Extracting Textual Content: We first extract text from the input NIPF document collections and then filter/clean-up the extracted text. The NIPF text input to Jaguar comes from all possible document types, including MS Word, PDF and HTML web pages, and is therefore prone to having many irregularities, such as incomplete, strangely formatted sentences, headings, and tabular information. The text extraction and filtering mechanism of Jaguar is a crucial step that makes the input acceptable for subsequent NLP tools to process it. The extraction/filtering rules include, conversion/removal of non-ASCII characters, verbalization of Wikipedia infoboxes and tables, conversion of punctuation symbols, among others.

2) Initial Seed Set Selection: For each NIPF topic, Jaguar is provided with an initial seed set containing on average 51 concepts of interest. The seed set is used to determine the set of text sentences of interest in a topic's document collection. The initial seed set selection for the NIPF topic was performed manually based on the concepts found in the topic descriptions. The initial seed selection process is the *only manual step* that we use in our NIPF ontology creation process. We are currently exploring automated methods for creating the initial seed set using a combination of statistical and semantic clues in the document collection.

3) Concept and Relation Discovery: For each NIPF topic, the set of text files extracted from the document collection are processed through the entire set NLP tools listed in Section II. The NLP processed data files are then passed through the concept discovery module, which identifies noun concepts in sentences which are related to the NIPF topic target words or seeds. The concept discovery module analyzes the syntactic parse tree of each processed sentence and scans them for noun phrases. Though Jaguar has the capability to extract verb concepts by analyzing verb phrases, for our current NIPF ontology creation experiment, we focused only on noun concepts and their semantic relations. Each noun phrase is then processed and well-formed noun concepts are extracted based on a set of syntactic patterns and rules.

Noun concepts (which are part of the seed set), their semantic relations (extracted from the semantic parser, Polaris [8], [9]) and the noun concepts involved in semantic relations with the seed set concepts are added into data structures for subsequent processing into the ontology's hierarchy. The resulting data structures are processed and used to populate one or many semantic contexts, groups of relations or nested contexts which hold true around a common central concept. The seed set is then augmented with concepts that have hierarchical relations with the target words or seeds. The entire process of sentence selection, concept extraction, semantic relation extraction and seed concepts set augmentation is repeated in an iterative manner, n number of times (by default, n is set to 3). While processing the NIPF topic collections through Jaguar, we used ISA, Part-Whole and Synonymy semantic relations for automatically augmenting the seeds concept set. Figure 2 depicts this iterative process of extracting concepts and semantic relations of interest using seed concepts.

4) Creating Concept Hierarchies: The extracted NIPF topic noun concepts and semantic relations are fed to the classification module to determine the hierarchical structure. Certain hypernymy relations discovered via classification contain anomalies (causing cycles) or redundancies. Hence, we run them through a *conflict resolution engine* to detect and correct inconsistencies. The *conflict resolution engine* creates a NIPF topic hierarchy link by link (relation by relation) and follows a conflict avoidance technique, wherein each new link is tested for causing inconsistencies before being added to the hierarchy.

5) Ontology Merging: Although single runs of Jaguar yield rich NIPF ontologies, Jaguar's real power lies in providing an ontology maintenance option to layer ontologies from many different runs. Figure 3 depicts the process of merging two ontologies through conflict resolution algorithms. Jaguar can merge disparate ontologies or add new knowledge by using the aforementioned conflict resolution techniques. The merge tool merges the two ontologies' concept sets, hierarchies (using conflict resolution), and their knowledge bases (set of semantic contexts). Given two ontologies or knowledge bases, ontology merging is performed by enumerating the relations in the smaller ontology and adding them to the larger or reference ontology. A relation may either be represented by a similar relation in the reference ontology, may create a redundant path between concepts or may be a new relation that can be added to the reference ontology. The conflict resolution techniques are then used for handling the conflict induced in the ontology to generate a merged ontology. Merging is useful for distributed or parallel systems where small chunks of the input text may be processed on some portions of the system and then subsequently merged. It also provides a foundation for future work in contextual reasoning and epistemic logic. The resulting rich NIPF knowledge bases can be viewed at many different levels of granularity, providing an analyst with the level of detail desired.

III. Evaluation of Jaguar's NIPF Ontologies

Since the mid-1990s, various methodologies have been proposed to evaluate ontology generation/maintenance/reuse techniques [11]. All the proposed methodologies have focused

TABLE I
Subset of semantic relations used to evaluate the performance of Jaguar's automatic NIPF topical ontology generation from text.

Semantic Relation	Definition	Example	Code
ISA	X is a (kind of) Y	[XY] [John] is a [person]	ISA
Part-Whole/Meronymy	X is a part of Y	[XY] [The engine] is the most important part of [the car] [XY] [steel][cage] [YX] [faculty] [professor] [XY] [door] of the [car]	PW
Cause	X causes Y	[XY] [Drinking] causes [accidents]	CAU

TABLE II
Performance results for Jaguar's automatic topical NIPF ontology generation from text with respect to the semantic relations defined in Table I.

Number of Annotators	NIPF Topic	Precision		Coverage		F-Measure	
		Correctness	Correctness+ Relevance	Correctness	Correctness+ Relevance	Correctness	Correctness+ Relevance
3	Weapons	**0.610090**	**0.501499**	0.702424	0.657122	**0.653009**	0.568859
1	Missiles	0.533867	0.485364	0.793775	**0.777747**	0.63838	**0.597715**
2	Illicit Drugs	0.471938	0.274506	0.801422	0.701122	0.594053	0.39454
1	Terrorism	0.388788	0.291019	**0.822285**	0.776206	0.527953	0.423323

TABLE III
Semantic relation and concept extraction statistics for the evaluated NIPF ontologies presented in Table II.

NIPF Topic	Unique Semantic Relations					Unique Concepts		
	ISA	PW	CAU	Others	**Total**	In ISA/PW/CAU	Others	**Total**
Weapons	1683	766	113	946	3508	2620	1012	3473
Missiles	2939	2296	646	2692	8573	5982	3539	7873
Illicit Drugs	2356	2040	817	5464	10677	5107	4982	7935
Terrorism	2590	4219	1497	5405	13711	7929	6247	11638

on some facet of the ontology generation problem, and depend on the type of ontology being created/maintained and the purpose of the ontology [12]. It is noted that not much progress has been achieved in developing a comprehensive and global technique for evaluating the correctness and relevance of ontologies [13].

$$\begin{aligned}
&Pr(Correctness)=\frac{N_j(correct)+N_j(irrelevant)}{N_j(correct)+N_j(incorrect)+N_j(irrelevant)}\\
&Pr\begin{pmatrix}Correctness\\+\\Relevance\end{pmatrix}=\frac{N_j(correct)}{N_j(correct)+N_j(incorrect)+N_j(irrelevant)}\\
&Cvg(Correctness)=\frac{N_j(correct)+N_j(irrelevant)}{N_g(correct)+N_g(irrelevant)+N_g(added)}\\
&Cvg\begin{pmatrix}Correctness\\+\\Relevance\end{pmatrix}=\frac{N_j(correct)}{N_g(correct)+N_g(added)}
\end{aligned} \tag{1}$$

We evaluated the quality of Jaguar's NIPF ontologies by comparing them against manual gold annotations. Following the ontology evaluation levels defined in [12], our evaluations are focused on the *Lexical, Vocabulary, or Data Layer* and the *Other Semantic Relations* levels. For a NIPF topic, the ontology and document collection were manually annotated by several human annotators and used in the evaluation of the ontology. Viewing an ontology as a set of semantic relations between two concepts, the annotators:

- Labeled an entry *correct* if the concepts and the semantic relation are correctly detected by the system else marked the entry as *Incorrect*
- Labeled a *correct* entry as *irrelevant* if any of the concepts or the semantic relation are irrelevant to the domain
- From the sentences *added new entries* if the concepts and the semantic relation were omitted by Jaguar

The annotation rules provide feedback on the automated concept tagging and semantic relation extraction and also are used for computing precision (Pr) and coverage (Cvg) metrics for the automatically generated ontologies. Equations in (1) capture the metrics defined by Lymba to evaluate Jaguar's automatic topical NIPF ontology generation from text. In (1), $N_j(.)$ gives the counts from Jaguar's output and $N_g(.)$ correspond to counts in the user annotations. Table II presents our initial evaluation results for 4 NIPF topics using a subset of 3 semantic relations (ISA, PW and CAU relations) defined in Table I. Table III presents the semantic relation and concept extraction statistics for the four NIPF ontologies being evaluated in this paper.

We use the metrics defined in (1) to evaluate the ontologies against the manual annotations from different human annotators. The results in Table II represent the evaluation scores which have been averaged over the results for different annotators. The first column in Table II identifies the number of annotators for each topic. Jaguar obtained the best *Precision* results in both *Correctness* and *Correctness+Relevance* evaluations for the *Weapons* NIPF topic. Please note that as shown in Table III, smaller number of concepts/semantic-relations were extracted for this topic due to its smaller collection size (50 documents versus the 500 document set for the other topics). The *Terrorism* NIPF topic obtained the best *Coverage* result for the *Correctness* evaluation and it was also very close to the best *Coverage* result obtained by the *Missiles* NIPF topic for the *Correctness+Relevance* evaluation. The *Weapons* NIPF topic obtained the best *F-Measure* result ($\beta = 1$) for the *Correctness* evaluation while the *Missiles* NIPF topic obtained the best *F-Measure* result for the *Correctness+Relevance* evaluation.

IV. Conclusions and Future Work

In this paper, we presented the semi-automatic development of an ontology library for the NIPF topics. We use Jaguar-KAT, a state-of-the-art tool for knowledge acquisition and domain understanding, with minimized manual intervention to create NIPF ontologies loaded with rich semantic content. We also defined evaluation metrics to assess the quality of the NIPF ontologies created using our methodology. We evaluated a subset of Jaguar's NIPF ontologies by comparing them against manual gold annotations. The results look very promising and show that a decent amount of knowledge was automatically and accurately extracted by Jaguar from the input document collection while keeping the manual intervention in the process to a minimum. We plan to perform further analysis of the results and identify methods for improving the precision and coverage of text processing and ontology generation.

References

[1] D. Bixler, D. Moldovan, and A. Fowler, "Using knowledge extraction and maintenance techniques to enhance analytical performance," in *Proceedings of International Conference on Intelligence Analysis*, 2005.

[2] P. Cimiano, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, 2006.

[3] D. Moldovan, M. Srikanth, and A. Badulescu, "Synergist: Topic and user knowledge bases from textual sources for collaborative intelligence analysis," in *CASE PI Conference*, 2007.

[4] E. Ratsch, J. Schultz, J. Saric, P. C. Lavin, U. Wittig, U. Reyle, and I. Rojas, "Developing a protein-interactions ontology," *Comparative and Functional Genomics*, vol. 4, no. 1, pp. 85–89, 2003.

[5] H. Pinto and J. Martins, "Ontologies: How can they be built?" *Knowlegde and Information Systems*, vol. 6, no. 4, pp. 441–464, 2004.

[6] "FBI: National Security Branch - FAQ," Last accessed on Jul 21, 2008, available at http://www.fbi.gov/hq/nsb/nsb_faq.htm#NIPF.

[7] D. I. Moldovan and R. Girju, "An interactive tool for the rapid development of knowledge bases," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 1-2, pp. 65–86, 2001.

[8] A. Badulescu, "Classification of semantic relations between nouns," Ph.D. dissertation, The University of Texas at Dallas, 2004.

[9] R. Girju, A. M. Giuglea, M. Olteanu, O. Fortu, O. Bolohan, and D. Moldovan, "Support vector machines applied to the classification of semantic relations in nominalized noun phrases," in *Lexical Semantics Workshop in Human Language Technology (HLT)*, 2004.

[10] G. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[11] Y. Sure, G. A. Perez, W. Daelemans, M. L. Reinberger, N. Guarino, and N. F. Noy, "Why evaluate ontology technologies? because it works!" *IEEE Intelligent Systems*, vol. 19, no. 4, pp. 74–81, 2004.

[12] J. Brank, M. Grobelnik, and D. Mladenic, "A survey of ontology evaluation techniques," in *Data Mining and Data Warehouses (SiKDD)*, Ljubljana, Slovenia, 2005.

[13] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann, "Modelling ontology evaluation and validation," in *European Semantic Web Symposium/Conference (ESWC)*, 2006, pp. 140–154.