

A RISK ASSESSMENT SYSTEM WITH AUTOMATIC EXTRACTION OF EVENT TYPES

Philippe Capet¹, Thomas Delavallade¹, Takuya Nakamura², Agnes Sandor³,
Cedric Tarsitano³, Stavroula Voyatzi²

¹*THALES Land & Joint Systems*
160 boulevard de Valmy, 92704 Cedex
first_name.last_name@fr.thalesgroup.com

²*Universite de Marne-la-Vallee, Institut Gaspard-Monge*
5 bd Descartes, 77454 Marne-la-Vallee Cedex 2
first_name.last_name@univ-mlv.fr

³*Xerox Research Centre Europe*
6 chemin de Maupertuis, 38240 Meylan, France
first_name.last_name@xrce.xerox.com

Abstract In this article we describe the joint effort of experts in linguistics, information extraction and risk assessment to integrate EventSpotter, an automatic event extraction engine, into ADAC, an automated early warning system. By detecting as early as possible weak signals of emerging risks ADAC provides a dynamic synthetic picture of situations involving risk. The ADAC system calculates risk on the basis of fuzzy logic rules operated on a template graph whose leaves are event types. EventSpotter is based on a general purpose natural language dependency parser, XIP, enhanced with domain-specific lexical resources (Lexicon-Grammar). Its role is to automatically feed the leaves with input data.

1. Introduction

In various fields rational risk analysis is part of the decision making process. It is a fundamental methodological tool which helps economic and political actors to anticipate potential crises. Such an analysis is usually carried out by human experts. The first step in risk analysis is the retrieval of relevant information from available data. The amount of the data may be so large that there is a great need for tools that automate parts of the risk analysis. An early

warning system should help experts to monitor massive flows of events, in the short term, and launch alerts whenever critical event sequences are detected.

For this purpose we are designing ADAC [7], an automated early warning system that provides a dynamic synthetic picture of situations involving risk. ADAC is being implemented for detecting weak signals of nuclear proliferation, in order to issue alerts about emerging nuclear risks as early as possible. Risk assessment in this domain has to process large amounts of knowledge -such as educational changes in a particular country, public statements of local leaders, covert information, diplomatic negotiations, satellite observations, etc- that can only be acquired through widely disparate channels of information. A significant amount of this knowledge is directly derivable from the events described in information newswires. At present, data concerning events are introduced into the ADAC system by human analysts.

However, the exponentially growing information flow through the internet no longer allows human analysts to keep abreast of the events referred to in the newswire sources. On the other hand, the more extensively a risk assessment system is populated the more reliable it is. Thus the use of an automatic information extraction (IE) system has become a necessary component of any risk assessment system based on the continuous monitoring of event flows.

In this article we describe the underlying principles of ongoing work: the joint effort of experts in linguistics, IE and risk assessment to integrate EventSpotter, an automatic event extraction engine, into ADAC.

This paper is organized as follows: In section 2 we describe ADAC, the risk assessment component that needs to be fed with automatically extracted events. In section 3 we present EventSpotter pointing out its innovative features compared to other event extraction systems. We argue that these features are necessary in order to meet the requirements of the subsequent risk assessment modules. In this section we underline the importance of the integration of extensive lexical resources into the event extraction system, and briefly describe their form and the principles that lead us to constitute them. We also present an evaluation of the present state of the IE system. In section 4 we present some related work in IE applied to event extraction. Finally, in section 5 we draw some conclusions and show directions for future work.

2. The Risk Assessment System: ADAC

ADAC is a dynamic risk assessment system that monitors the daily evolution of the situation in a particular domain in order to give experts a better understanding of situations involving risk. Based on a library of experts' scenarios describing typical crisis developments and on an ontology representing the domain knowledge, the system monitors a flow of incoming event in order

to spot the event sequences that are likely to end up in crisis. This section is dedicated to the presentation of the various components of the ADAC system.

2.1 Scenarios

A scenario describes typical developments of a specific type of crisis, i.e. it depicts how a system moves from a normal, sound state to a critical, anarchical state. Scenarios are expressed in the template formalism. The general principle of a template is the description of a complex phenomenon as a combination (conjunction, disjunction, etc.) of less complex phenomena which again are decomposed into a combination of less complex phenomena until elementary phenomena are reached, namely the events directly observed from the input.

2.2 Ontology

The scenarios, represented by templates, are part of the knowledge base feeding the system, which is more general than a simple scenario library. It gathers all the information the system has at its disposal to perform crisis detection. It contains all the linguistic terms, organized in a subsumption hierarchy, that are necessary to link the scenarios with the input data. Locations and actors that are under watch are for instance defined in this ontology.

2.3 Event Data

The data used to feed the system are structured representations of occurring events, related to the particular type of crisis under study. The structure used to summarize a piece of information is defined by the following fields: the source reporting the information, the date and location of the reported event, the event type, the actors of the event (persons or organizations), the other persons or organizations involved in the event (other participants), the uncertainty of the event from the source's point of view (does the source report the event as a fact, is it an assumption, an opinion...?). The choice of these fields has been inspired by the work of political scientists concerning the anticipation, monitoring and termination of wars [14].

2.4 Recognition Engine

The core of our system consists in comparing input event data and known scenarios of crisis developments, through a constrained pattern matching process. This is done by our recognition engine, which assesses the degree of match between event sequences and experts' scenarios, taking into account some, spatial, temporal and operational constraints. It may indeed be important to ensure that two events are considered as parts of the same scenario, only if they occur in a specific area, within a given time frame, while involving sim-

ilar actors. The recognition degree is estimated by a similarity degree, which takes its values between 0 (null recognition) and 1 (full recognition). It evolves according to the arrival of new events which confirm or invalidate the hypothesis: each time new data match the template, the recognition degree is updated. Through this information fusion mechanism, guided by experts' knowledge, ADAC enables to detect the early signs of what usually ends up in crisis.

3. Automatic Extraction of Event Types: EventSpotter

3.1 General Properties

As we outlined above, the ADAC system calculates risk on the basis of a template graph whose leaves are event types. The role of the natural language processing (NLP) system is to feed the leaves with input data. We do this by extracting event descriptions from sentences of news articles. These extracted event descriptions are transformed by subsequent operations, which render them suitable for further automatic processing in the template.

The extraction of event descriptions is carried out with syntactic analysis using the Xerox Incremental Parser (XIP) [2]. An event is defined in terms of syntactic relationships in the sentences. Out of all the dependency relations produced by the analyzer, we consider an event description to be a predicate (verb, adjective and predicative noun) related to its arguments and modifiers.

The first operation that EventSpotter carries out with respect to the extracted event descriptions is their normalization to a unique representation structure. This operation is domain-independent since invariably every event description is extracted. The unique representation structure consists first in indicating for each event description its information source and factuality as conveyed by the information source. Furthermore we transform each event description into a set of common constituents. The constituents of events are a core, which is the name of the event, and its coordinates, whenever they are present in the sentence. We have defined the coordinates of event cores as agent(s), other participant(s), place(s) and time. The way this normalization is carried out by EventSpotter is described in a previous article [12].

The second operation that EventSpotter carries out is the association of the extracted event descriptions to the pre-defined relevant event types that constitute the leaves of the template graph in ADAC, whenever it is appropriate. This operation is domain-specific. It is carried out on the event core and its extensions, as we will describe below, and as it is illustrated by the following example taken from the corpus that we have chosen for developing our application. Sentences (2) through (4) indicate extended event cores in bold. These extended event cores are the parts of the sentences associated with (1), one of the relevant event types with respect to our domain defined in ADAC.

(1) to get involved in a cooperation in the nuclear domain.

- (2) A delegation from Syria arrives in Iran **to begin negotiations on a possible Iranian-Syrian nuclear pact**.
- (3) The Middle East Newslines reports that Iran **is preparing to receive a light water nuclear reactor** from Russia.
- (4) Former chief nuclear negotiator for Iran Hassan Rowhani says Tehran **is ready to negotiate a mutual start for the Natanz nuclear facility**.

After the second operation, Table 1 is extracted.

Table 1. Representation of the extracted event (3)

<i>Source</i>	<i>Fact.</i>	<i>Actor</i>	<i>Core</i>	<i>Oth.pt</i>	<i>Place</i>	<i>Time</i>	<i>Event type</i>
Middle East Newslines	F	Iran	receive a light water nuclear reactor	Russia			to get involved in a cooperation in the nuclear domain

3.2 Concept-Matching

In order to match event descriptions in sentences (2) through (4) with the relevant event type (1) we use the concept-matching framework. Since we have described it previously [13], here we only recall the basic idea. Concept matching combines the bag-of-words approach with syntactic dependency parsing for extracting complex target concepts. The complex target concepts are coherent and recurrent meaning fragments of sentences, and are expressed in highly diverse ways. Within the concept-matching framework the target concepts (like (1) above) are matched whenever syntactically related chains of expressions conveying - what we call - their constituent concepts (bags of words) occur within the same sentence. We show how the concept-matching framework is applied on (2) through (4) for matching the target concept (1).

As first step, the target concept is broken down into three constituent concepts: [get involved] in [cooperation] in the [nuclear domain]. We assign general concept labels to these constituent concepts as follows: BEGIN LINK NUCLEAR.

The sentences (2) through (4) all contain words that convey each of these concepts. Moreover, these words form dependency chains in all of the sentences as shown below, thus they do in fact convey the target concept (1):

- (5) A delegation from Syria arrives in Iran to <BEGIN> begin </BEGIN>
 <LINK> negotiations </LINK> on a possible Iranian-Syrian
 <NUCLEAR> nuclear pact </NUCLEAR>.

dependency chain: ... begin ... negotiations on ... nuclear pact.

- (6) The Middle East Newslines reports that Iran <BEGIN> is preparing </BEGIN> <LINK>to receive</LINK> a light water <NUCLEAR> nuclear reactor </NUCLEAR> from Russia.

dependency chain: ... is preparing to receive ... nuclear reactor ...

- (7) Former chief nuclear negotiator for Iran Hassan Rowhani says Tehran <BEGIN> is ready </BEGIN> <LINK> to negotiate </LINK> a mutual <BEGIN> start </BEGIN> for the Natanz <NUCLEAR> nuclear facility </NUCLEAR>.

dependency chain: ... is ready to negotiate a ... start for ... nuclear facility.

We can observe that neither the order nor the type of relationship among the constituent concepts is relevant for the match. The essential constraint of coherence is the existence of a dependency relationship among the words conveying the constituent concepts.

In order to carry out concept-matching automatically we need a general purpose natural language dependency parser as well as domain-specific resources: a set of constituent concepts, lists of words conveying the constituent concepts, a lexicon-grammar to improve the performance of the general purpose parser in the establishment of the argument and modifier dependencies and rules of the co-occurrence of the constituent concepts. The general purpose dependency parser is used for extracting all the possible dependency pairs among the words of the sentences, whereas the domain-specific resources make it possible that the relevant dependency pairs conveying the constituent concepts can be chosen out of all the dependency pairs in the sentences and associated with the target concepts.

In the following sections we will concentrate on the domain-specific resources we have used to extract descriptions of event types that are needed by ADAC to calculate risks of nuclear proliferation.

The Target Concepts and the Constituent Concepts for Nuclear Proliferation. In ADAC 103 event types are listed as relevant for inducing crisis in the nuclear domain. Table 2 is an excerpt of the list.

Table 2.

<ol style="list-style-type: none"> 1. to work on secret nuclear programs 2. to sell military equipment 3. to get involved in a cooperation in the nuclear domain

In the entire list we propose 18 constituent concepts, whose various co-occurrence combinations cover all the target concepts. They are the following:

NEGATIVE, INTENT, BEGIN, CONTINUE, END, POSITIVE, HOSTILITY, LEGAL, SECRET, MILITARY, NUCLEAR, KNOWLEDGE, LINK, MOVEMENT, PRODUCTION, MONEY, TOOL, STATE

The granularity of the constituent concepts is subject to experimentation. If certain constituent concepts do not assure fine-grained event types, they can be broken down to several types. A word might also be assigned to several constituent concepts. Table 3 shows Table 2 marked up with constituent concepts:

Table 3.

1. to work[PRODUCTION] on secret[SECRET] nuclear[NUCLEAR] programs[PRODUCTION]
2. to sell[MONEY] military[MILITARY] equipment[TOOL]
3. to get involved[BEGIN] in a cooperation[LINK] in the nuclear domain[NUCLEAR]

For the present system we have assigned the list of words to the constituent concepts manually. We have worked on a prototype based on a corpus of news containing 4196 tokens of content words.

Table 4. Some sample constituent concepts and some words associated with them

Constituent concept	Words
NEGATIVE	contrary, lie, refute
INTENT	decide, effort, require
LINK	ally, connect, negotiate
CONTINUE	augment, emerge, regular

Lexicon Grammars. As we pointed out above we have built domain-specific lexical resources in order to ensure the precision of the extraction of the relevant dependencies in the sentences. In the general-purpose XIP parser certain dependencies, especially prepositional phrase attachment and clausal complementation cannot be handled with high precision due to word sense ambiguities on the one hand and the lack of a broad coverage lexical grammar on the other hand. Working on a domain-specific vocabulary allows us to build lexicon-grammars for this vocabulary.

A lexicon-grammar is a dictionary that provides exhaustive and detailed subcategorisation information about the predicates of a natural language such as verbs, predicative nouns and adjectives. Predicates with related syntactic and semantic behaviour are grouped together, for example, in the structure of simple sentences, in the distribution of arguments and in terms of interpretations [10]. The lexical, syntactic and semantic features provided by the lexicon-grammars are used for establishing grammatical and dependency rules. Table 5 shows an extract of the lexicon-grammar of English verbs taking a sen-

tential complement (e.g. *Russia admitted (that Iran's program is of peaceful intent + having discussions with Iran + to providing incorrect information)*).

Table 5.

	NO=: Nhum	NO=: N=hum	NO=: Nsupport	NO=: Npl obl		NO V	NO V NI	NO V with N2hum NI	NI=: that S	NI=: that Ssubj	NI=: Wh-S	NI=: if S	NI=: whether S	NI=: VOing W	NI=: to VOing W	NI=: on VOing W
+	-	+	-		acknowledge	-	+	-	+	-	-	-	-	+	-	-
+	-	+	-		add	-	+	-	+	-	-	-	-	-	-	-
+	-	-	-		admit	-	+	-	+	-	+	+	+	+	+	+
+	-	-	+		agree	-	+	+	+	-	-	-	-	-	-	+

Sentence (8) is associated to the event-type "nuclear-related agreement" due to specific syntactic features present in the lexicon-grammar, which allow XIP to extract a dependency between "agreement" and "import":

- (8) An unnamed official of the Russian atomic energy ministry says that Russia has yet to receive Iran's **agreement** for Moscow **to import back radioactive fuel waste** from an Iranian nuclear power plant that Russia is building in Bushehr.

dependency chain: ... agreement ... to import back ... radioactive fuel waste

We carried out an evaluation of in what extent the lexicon-grammar resources have influenced the extraction of dependency relationships. We established a gold standard of a 100 sentences, which we compared to the output of XIP with and without the addition of the lexicon-grammar. The improvement of the performance was 36%, which is a significant difference.

4. Related Work

Event extraction is the subject of an increasing number of information extraction applications. Different systems, however, represent events in different ways. [1] describes two approaches to represent events: "On the one hand, there is the TimeML model, in which an event is a word that points to a node in a network of temporal relations. On the other hand, there is the ACE model, in which an event is a complex structure, relating arguments that are themselves complex structures, but with only ancillary temporal information.". Our representation is closer to the Automatic Content Extraction (ACE) model, which however, does not describe it entirely. We are not aware of any system that shares our event representation.

Apart from differences of representation, event extraction has been handled with various approaches. It is difficult to make comparisons among differ-

ent systems with different purposes in this article. We will give here a brief overview of the latest systems whose main application is event extraction.

Recent applications in the field of event extraction have mainly been carried out with probabilistic or machine learning approaches, which do not need to rely on strict linguistic constraints, but which fail to extract exact semantic relationships based on sentence structure. The following articles describe various event extraction systems, and none of them shares our approach and purposes: [11] built a Retrospective news Event Detection system which merges events with existing similar events with a probabilistic approach based on bag-of-words and clustering. [15] built a prize-winning event extraction system based on machine learning with limited linguistic constraints. [4] study the tradeoffs between open and traditional relation extraction. They conclude that it seems more interesting to use traditional IE for a domain specific extraction. Finally, closer to our approach, [3] built a relation and event extraction system, but only for verb-based events.

Several works in NLP systems argue that acquiring subcategorization information is an important task for the improvement of performance (see [5; 6; 8; 9]). Some of them also put forth that manual acquisition of such resources is time and resource consuming (see [6]). However, manually-developed lexicons (enriched with subcategorization information) prove to be precise [6]. Moreover, [5] estimate that half of the parse failures is caused by inaccurate or incomplete subcategorization information [8].

5. Conclusion

In this article we have described the components of an entirely automatic integrated system of risk assessment concerning nuclear proliferation. It consists basically of two components: linguistic analysis of news articles and computation of nuclear risk. The actual integration has not taken place technically but the evaluation of the output of the linguistic component shows that conceptually it is possible, i.e. the two components are compatible. Once the integration is carried out, ours will be the first system where automatic linguistic analysis of newswire articles is used as input to a risk detection system.

This system is the result of a chain of processes where in each step the enrichment of a lower level analysis makes way to a higher level analysis. Starting from lexical analysis, continuing by syntactic parsing of free text, the output of which is then mapped into semantic role assignment, and coupled with conceptual analysis, the process goes on by carrying out operations on high-level concepts, which yields the final output.

Further work consists in the integration of the two components, which will be followed by an evaluation. We carried out a partial evaluation of the system, which we reported in previous work [12]. In a longer term perspective, the

approach that we propose for nuclear risk assessment can be extended for other kinds of political risk assessment or in general to any kind of risk assessment where the risk is related to event occurrences that are reported in texts.

Acknowledgments

This research is being funded by the French national project Infom@gic.

References

- [1] Ahn, D.: *The stages of event extraction*. In: Proceedings of the Workshop on Annotations and Reasoning about Time and Events, pp. 1–8. (2006)
- [2] Ait-Mokhtar, S., Chanod, J-P., Roux, C.: *Robustness beyond shallowness: incremental dependency parsing*. Natural Language Engineering, 8(2/3) pp. 121–144. (2002)
- [3] Aone, C., Ramos-Santacruz, M.: *REES: A Large-Scale Relation and Event Extraction System*. In: Proceedings of the sixth conference on Applied natural language processing, pp. 76–83. Seattle, Washington (2000)
- [4] Banko, M., Etzioni, O.: *The Tradeoffs Between Open and Traditional Relation Extraction*. ACL (2008)
- [5] Briscoe, T., Carroll, J.: *Generalised Probabilistic LR Parsing for Unification-Based Grammars*. Computational Linguistics, 19(1) (1993)
- [6] Carroll, J., Fang, A.: *The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser*. In: Proceedings of the First International Joint Conference on Natural Language Processing, pp. 107–114. Sanya City (2004)
- [7] Delavallade, T., Mouillet, L., Bouchon-Meunier, B., Collain, E.: *Monitoring Event Flows and Modelling Scenarios for Crisis Prediction: Application to Ethinc Conflict Forecasting*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. (2007)
- [8] Gardent, C., Guillaume, B., Falk, I., Perrier, G.: *Le lexique-grammaire de M. Gross et le traitement automatique des langues*. In ATALA (2005)
- [9] Korhonen, A.: *Semantically Motivated Subcategorization Acquisition*. In: Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition, 9, pp. 51–58. Philadelphia (2002)
- [10] Leclere, C.: *Organization of the Lexicon-Grammar of French Verbs*. Lingvisticae Investigationes, 25(1), pp. 29–48 (2002)
- [11] Li, Z., Wang, B., Li, M., Ma, W-Y.: *A Probabilistic Model for Retrospective News Event Detection*. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 106–113. Salvador (2005)
- [12] Rebotier, A., Sandor, A., Voyatzi, S., Nakamura, T., Martineau, C., Delevallade, T., Capet, P., Jacquelinet, J.: *Intelligent awareness: event extraction, information evaluation & risk assessment*. In: 3rd Language & Technology Conference, pp. 539–543. Poznan (2007)
- [13] Sandor, A., Kaplan, A., Rondeau, G.: *Discourse and Citation Analysis with Concept-Matching*. In: International Symposium, Discourse and Document, pp. 147–151. Presse Universitaire de Caen, Caen (2006)
- [14] Schrodtt, P., Davis, S., Weddle, J.: *Political Science: KEDS-A Program for the Machine Coding of Event Data*. Social Science Computer Review. 12, 561–588 (1994)
- [15] Xu, F., Uszkoreit, H., Li, H.: *Automatic Event and Relation Detection with Seeds of Varying Complexity*. AAAI Workshop Event Extraction and Synthesis, Boston (2006)