



From the texts to the concepts they contain: a chain of linguistic treatments



Ahmed Amrani¹, Jérôme Azé², Thomas Heitz², Yves Kodratoff², Mathieu Roche²

¹ ESIEA Recherche, 9 rue Vésale, 75005 Paris - France
² Équipe Inférence et Apprentissage (IA) - Laboratoire de Recherche en Informatique (LRI) - Université Paris-Sud - Orsay – France
E-mail: {amrani,aze,heitz,yk,roche}@lri.fr

Abstract: The text-mining system we are building deals with the specific problem of identifying the instances of relevant concepts present in the texts. Therefore, our system relies on interaction between a field expert and the various linguistic modules we use, often adapted from existing ones, such as Brill's tagger or CMU's Link parser. We have developed learning procedures adapted to various steps of the linguistic treatment, mainly for grammatical tagging, terminology, and concept learning.

First step : Normalization

- Homogenization of the TREC Novelty corpus
- Creation of a lexicon

Second step : Tagging

- Grammatical tag : Brill's tagger (Brill, 1994) for the field.
- ETIQ + dedicated language

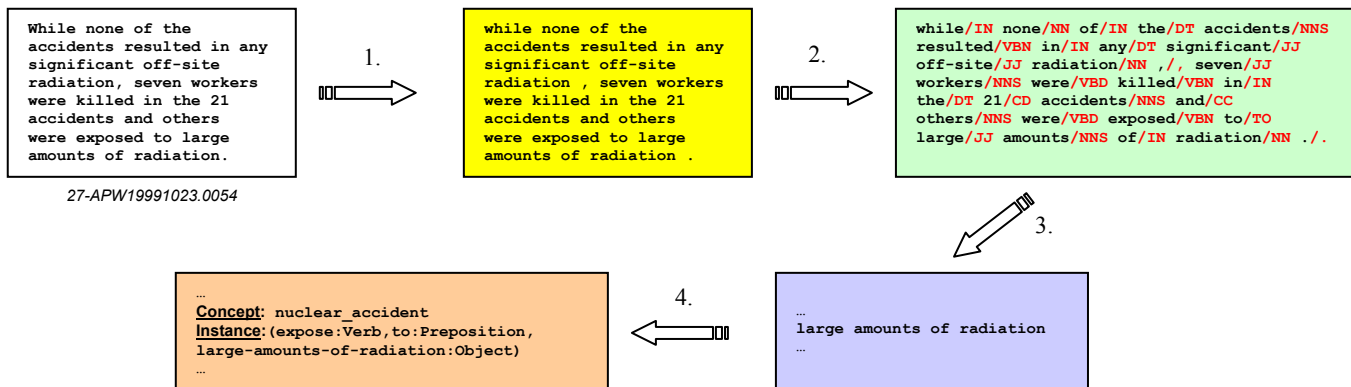
Third step : Terminology

- Iterative extraction of the more relevant terms (Adjective-Noun, Noun-Noun, etc.).
- EXIT **EX**tracts Iteratively complex Terms.

Fourth step : Taxonomy

Semi-automatic approach:

- Link Grammar parsing.
- Using ACT with an expert, (ACT : software developed by our team for the recognition of linguistic instances of concepts).



Task 1.1 : Detection of relevant sentences

Each sentence is replaced by a summary containing only

- individual present in the topic and/or texts
- terms and their various forms (find with FASTR)
- places names, numbers, verbs and noun present in the topic

- Automatic approach : compute for each sentence and select the ones with a score higher than the average score.
- Semi-Automatic approach : using the automatic approach, together with expert defined concepts.

Task 1.2 : Detection of novel sentences

A sentence is novel if it contain a novel information, i.e., new words or concepts.

Task 2 : Detection of novel sentences

Using the relevant sentences, we computed a score for each sentence. This score is based on the sum of the TF x IDF words of the sentences (except stop-words).

To determine if a sentence is novel or not, we used two threshold:

- a dynamic one based on the TF x IDF score
- a static one based on the percentage of novel information contained in the sentence

	Task 1.1	Task 1.2	
Run 1	0.306	0.066	Individual from topic and texts
Run 2	0.356	0.108	Individual from topic or texts
Run 3	0.255	0.098	10% Run 1, 90 % Concepts
Run 4	0.299	0.098	50 % Run 1, 50 % Concepts
Run 5	0.302	0.072	90% Run 1, 10% Concepts

Average Fscore on the five runs.

Run 1	0.614	no concepts, no coreferences, 5%
Run 2	0.614	no concepts, coreferences, 5%
Run 3	0.598	concepts, no coreferences, 5%
Run 4	0.597	concepts, coreferences, 5%
Run 5	0.602	concepts, coreferences, 20%

Average Fscore on the five runs.