
Bitext - Information Extraction

Bitext's information extraction tools provide all the capabilities available from language technologies to any application which manages large amounts of text:

- Business Intelligence
- Press Tracking
- Market Analysis
- Technology Watch
- Company reputation
- Fraud detection, etc.

Functionalities

These tools allow to extract various types of information from large textual databases:

- Entities: proper names of people, companies, products, events, etc.
 - Barack Obama, World Bank, US Government
- Concepts: ideas or topics in a piece of text
 - Global warming, developing countries, international cooperation
- Events or relationships between entities and concepts
 - <World Bank, increase, international cooperation>

These applications can be adapted to solve specific problems such as:

- to detect all the appointments of new positions which appear in various sources, like specialized press (for the private sector) or official bulletins (for the public sector)
- to create a list of all the physical and legal persons who establish civil relationships, commercial relations, etc., in legal documents, such as in deeds or bills of sales
- to collect all the press news about a specific company or topic

Moreover, these applications allow for the creation and maintenance of the knowledge bases (specialized dictionaries) which are specific for each customer, in order to tackle problems such as automatic cataloguing or ontology creation. These tools offer a wide variety of customization and configuration features.

These applications are available for Spanish, English, French, Catalan, Valencian, and Basque. In addition, the following languages are under development: Galician and Portuguese.

Therefore, Bitext increases its offer of language technologies solutions in the field of information extraction and content analysis (also known as "text analytics", "text mining", "content intelligence" or "business intelligence").

Next, various examples of entity, concept and event extraction are shown.



Examples of Entity Extraction

The entity extraction applications allow to analyze a text and get:

- place names (New York, Fifth Avenue), personal names (Carl Biden, Nancy Smith), names of organizations (IMF, Ministry of Health), etc.
- attributes of those people (judge, Field Marshal)
- relationships between people and organizations (Carl Biden chairs the IMF)

In addition, the application can link these entities to dictionaries, thesauri, ontologies and other knowledge sources in order to provide a complete context for a specific entity.

A fragment from a press article about the H1N1 virus is shown next:

The objective is to increase this figure up to 1200 million doses, so that the demand of the new seasonal influenza and H1N1 virus flu can be met. Another issue which needs to be addressed with pharmaceutical companies is the price of the treatment. The vaccine does not have to be expensive (seasonal vaccines can cost 20 euros per person), but nevertheless there will be countries which cannot afford them. The World Health Organization (WHO) claims that its objective is to get at least 110 million free doses from laboratories for the poorer or, alternatively, funds from the World Bank or from UN organizations which have the capacity to fund it: Unicef and the Pan American Health Organizations (PAHO), as Marie-Paule Kieny explained, who warned that all the plans can vary should the virus change. There is more information about the pathogen every day and not all the news are bad ones. The expansion seems to have slowed down, according to the latest data from WHO. Besides, experts get to better know the agent and its behavior. Centers for Disease Control from the USA announce that they have developed rapid diagnostic kits which have already been sent to all the States.

From the previous fragment, Bitext's entity extraction application will generate the following list of entities, highlighting whether or not they are present in the dictionary of the application:

DETECTED ENTITIES	PRESENT IN THE DICTIONARY	NOT PRESENT IN THE DICTIONARY
H1N1	-	YES
WHO	YES	-
World Bank	YES	-
UN	YES	-
Unicef	-	YES
Pan American Health Organization (PAHO)	-	YES
Marie-Paule Kieny	-	YES
WHO	YES	-
Centers for Disease Control from the USA	-	YES
States	-	YES

The application can also identify alternative forms of referring to the same entity (for example, it will suggest that "Kieny" is highly likely to be the same entity as "Marie-Paule Kieny").



Examples of Concept Extraction

The concept extraction applications analyze a text and extract concepts based on linguistic structures (noun phrases, prepositional phrases, etc.). Concept extraction also includes entity extraction.

As in the previous case, the application can categorize texts by linking concepts and dictionaries, thesauri, ontologies and other knowledge sources.

A fragment from the same press article about the H1N1 virus is shown next:

Opposite to the situation in Mexico, the flu epicenter, where according to the Health Secretary, José Ángel Córdovas, the rate of transmission is slowing down. US health authorities warned yesterday that they are expecting more cases, more hospitalizations and more deaths. The good news comes from a study of the data about the outbreaks in the United States, the country registering more cases after Mexico. One study by Ira Longini, epidemiologist from University of Washington, in Seattle, states that the infectivity rate (the number of people who get infected from an infected person) is 1.

From the previous fragment, Bitext's concept extraction application will extract the following list of concepts (entities in bold):

CONCEPTS - NOUN PHRASES
<i>Situation</i>
Mexico
<i>flu epicenter</i>
Health Secretary
José Ángel Córdovas
<i>rate of transmission</i>
US health authorities
<i>cases</i>
<i>hospitalizations</i>
<i>deaths</i>
<i>good news</i>
<i>study</i>
<i>data about the outbreaks in the United States</i>
<i>country</i>
<i>cases</i>
Mexico
<i>study</i>
Ira Longini
<i>epidemiologist from University of Washington</i>
Seattle
<i>infectivity rate</i>
<i>number of people</i>



Examples of Event Extraction

These event extraction applications analyze a text and extract not only entities and concepts, but also events, that is, facts or actions carried out by entities and concepts, for example

"José Pérez chaired the meeting of finance ministers"

- Event
 - Entity: José Pérez
 - Type of event: to chair
 - Concept: meeting of finance ministers

A fragment from the same press article about the H1N1 virus is shown next:

In Spain, the latest data from the Ministry of Health state that 9 out of 81 infected people have not been in Mexico. This way, the minister Trinidad Jiménez said yesterday in Zaragoza that the situation is very controlled. Meanwhile, the diplomatic war continues all over the world. The World Health Organization has sought explanations from the countries, like China, which have forbidden citizens of the affected States from entry. Spain continues the battle so that the veto of the pork products be lifted in Russia.

Bitext's event extraction application will extract the following list of analysis from the sentences in the fragment including entities, concepts and information about the who, what, how, when, etc.:

Sentence 1: "In Spain, the latest data from the Ministry of Health state that 9 out of 81 infected people have not been in Mexico."

- ⇒ **Entities:** Spain, Ministry of Health, Mexico
- ⇒ **Concepts:** latest data, 9 out of the 81 infected people
- ⇒ **Type of event:** to state
- ⇒ **Subject:** the latest data from the Ministry of Health
- ⇒ **Object:** that 9 out of the 81 infected people have not been in Mexico
 - **Type of event:** to be(negated)
 - **Subject:** 9 out of the 81 infected people
 - **Place:** Mexico
- ⇒ **Place:** Spain

Sentence 2: "This way, the minister Trinidad Jiménez said yesterday in Zaragoza that the situation is very controlled."

- ⇒ **Entities:** Trinidad Jiménez (minister), Zaragoza
- ⇒ **Concepts:** situation
- ⇒ **Type of event:** to say
- ⇒ **Subject:** the minister Trinidad Jiménez
- ⇒ **Object:** that the situation is very controlled
 - **Type of event:** to be
 - **Subject:** the situation
 - **Qualifying:** very controlled
- ⇒ **Manner:** this way
- ⇒ **Time:** yesterday

⇒ **Place:** Zaragoza

Sentence 3: "Meanwhile, the diplomatic war continues all over the world."

- ⇒ **Entities:** ---
- ⇒ **Concepts:** diplomatic war, world
- ⇒ **Type of event:** to continue
- ⇒ **Subject:** the diplomatic war
- ⇒ **Time:** meanwhile
- ⇒ **Place:** all over the world

Sentence 4: "The World Health Organization has sought explanations from the countries, like China, which have forbidden citizens of the affected States from entry."

- ⇒ **Entities:** World Health Organization, China, States
- ⇒ **Concepts:** explanations, citizens of the affected States, entry
- ⇒ **Type of event:** to seek
- ⇒ **Subject:** World Health Organization
- ⇒ **Object:** explanations
- ⇒ **Patient:** countries, like China, which have forbidden citizens of the affected States from entry
 - **Example:** China
 - **Type of event:** to prohibit
 - **Subject:** countries like China
 - **Object:** citizens of the affected States from entry

Sentence 5: "Spain continues the battle so that the veto of the pork products be lifted in Russia."

- ⇒ **Entities:** Spain, Russia
- ⇒ **Concepts:** battle, veto, pork products
- ⇒ **Type of event:** to continue
- ⇒ **Subject:** Spain
- ⇒ **Object:** battle
- ⇒ **Purpose:** so that the veto of the pork products be lifted in Russia
 - **Type of event:** to lift
 - **Subject:** the veto of the pork products
 - **Object:** ---
 - **Patient:** ---
 - **Place:** Russia

Bitext
The Bits and Text Company
Edificio Prisma, 1-1A - Cólquide 6
28230 - Las Rozas (Madrid) - SPAIN
<http://www.bitext.com>
info@bitext.com

