Uncertainty Detection For Information Extraction

Bénédicte Goujon
Thales Research & Technology
Campus de Polytechnique
1, avenue Augustin Fresnel
91 767 Palaiseau - FRANCE
benedicte.goujon@thalesgroup.com

Abstract

The detection of factual information from texts, like relations or events, is an important task in natural language processing. However most of the tools dealing with information extraction do not take into account the nuances expressed by the author, such as uncertainty. In those applications, the sentence "According to a witness, Y has met X" is reduced to its second part. The lost of the uncertainty expressed by the use of the secondary source "a witness" transforms the understanding of the whole information. The method presented in this paper aims at detecting the uncertainty and the reality of the information described in texts, and captures whether an information is presented as past, with a moderate uncertainty or in a negative way. The method is implemented in a web service for event annotation from texts, which is used in strategic watch applications.

Keywords

Semantic Annotation, Information Extraction, Event Extraction, Uncertainty, Linguistic patterns.

1. Introduction

The automatic extraction of events from texts has to take into account the discursive context in which the events are presented. For example, in the following sentences : "According to a witness, Y has met X", "Y may have met X", "X has met Y", several different levels of certainty are expressed by the author. In the first case, the author presents a reported discourse and the certainty of the event is relative to the reliability of the secondary source ("a witness" in this example). In the second case, the author expresses an uncertainty with "may". In the third case, the author expresses no uncertainty. Our aim is to capture those expressions of uncertainty, coupled with the expressions of reality of the information (for example, future tense is used when events have not occurred yet).

First, we present our context of information extraction. Then, we detail existing approaches in information extraction, modality and uncertainty detection and their limits. Afterwards, we present our model and the associated linguistic knowledge. We end with the implementation as web service and its evaluation.

2. Information Extraction

Our aim is to extract structured information from texts, in order to help the end-user like a strategic watch expert to identify a maximum of information for his or her task. Several difficulties are occurring. First, we want to extract all the information characteristics. For example if the end-user is interested in a Purchase event, we want to identify the agent, the patient, and according to the given details we want to identify all others parameters (date, location…). To do so, we need to build very precise linguistic patterns, if possible easy to build by the end-user. Another difficulty is the identification of all the nuances that are associated to the extracted information. For example, in the following sentence "Laurent Gbagbo might go to Italy.", the Move event relies Laurent Gbagbo and Italy, and is expressed with uncertainty. If this event is extracted and added into a knowledge base without this nuance, the result may not have the same sense as the original information in the text. Those slight differences, obvious when reading a text, must be taken into account by the automatic event extraction tools. The first needs in event extraction, expressed by the NIST MUC 3 (1991) and MUC 4 (1992) campaigns, were aiming at the identification of location, date and victim of past events [8]. They were not concerned by expression of uncertainty. More recently, the ACE (Automatic Content Extraction) campaigns integrated the event extraction, with the identification of attributes such as modality or polarity. However, in 2007, only one candidate (BBN Technologies) performed this test [1], which was suppressed in 2008. We may thus conclude that systems are not yet ready to such evaluations. The importance to manage the uncertainty expressed by the author of the text is now well identified. For example, Auger and Roy of the Defense R&D Canada [2] show the necessity to take into account the ambiguities and characterize automatically certainty/uncertainty expressed into texts in order to fuse information afterwards. Our aim is to identify the uncertainty related to events that are extracted.

3. State of the art and limitations

3.1 Information Extraction from Texts

Here is the presentation of two existing tools aiming at extracting events from texts.

Zenon [6] aims at extracting actions from HUMINT documents of the KFOR, in English. It uses FrameNet [4] to define the actions (KILL, REPORT, KNOW, COMMAND, PROPOSE, EXPLODE) and entities (Company, Person, Number, Date, City, Region, River, …)

International Conference RANLP 2009 - Borovets, Bulgaria, pages 118–122

to extract. Zenon is based on GATE [3], which is a free open source framework for Natural Language Engineering. This tool extracts actions and their corresponding entities but does not take into account nuances such as uncertainty that can modify the sense of the extracted actions.

The University College Dublin [7] has also developed a tool for the event extraction from heterogeneous sources. Their tool identifies sentences expressing events. Their aim is to cluster the sentences that express a same event, in order to ease the understanding of an event by a user. In this work, the objective is to extract complete sentences, but not to extract structured information from texts. So their approach keeps all the nuances of the author as it keeps the sentences, but it doesn't capture automatically events and their participants.

3.2 Modality and Event Detection

Saurí et al. [10] have worked on the identification of modality values associated to events described in texts. Their aim is to improve for example question-answering systems. Modalities taken into account in their approach are the following: degree of possibility, belief, evidentiality (*Subcomandante Marcos said that the Mexican government is not interested in putting an end to the conflict.*), expectation (*Hans Blix wants the US to allow UN inspectors back into Iraq to verify any weapons found by coalition forces.*), attempting and command. This work is based on the TimeML language. EvITA, a system using this approach, aims at recognize the events, and identifies among other things the modal characteristics. In this work, the certainty modality is not studied, whereas this modality is the most important for us.

3.3 Uncertainty Model for Natural Language

Several linguistic works aim at modeling the use of modality, but very few concentrate on uncertainty, for instance, the Certainty Categorization Model proposed by Rubin et al. in [9]. This model is based on four dimensions, called "level", "perspective", "focus" and "time", to characterize the uncertainty. Each of those dimensions are detailed in Figures 1 and 2 below, and may be illustrated with a few examples as follows.

Let us first consider the Level dimension: sentences (1) and (2) below give examples of an Absolute Level and a Low Level, respectively. (1) *Eventually, however, auditors* ***will almost certainly have to*** *form a tough self-regulatory body that can oversee its members' actions…*

(2) *So far the presidential candidates are more interested in talking about what a surplus* ***might buy*** *than in the painful choices that lie ahead.*

For the Perspective dimension, the example (3) illustrates a reported point of view.

(3) *The historian Ira Berlin, author of "Many Thousands Gone,"* ***estimates*** *that one slave perished for everyone who survived capture in the African interior…*

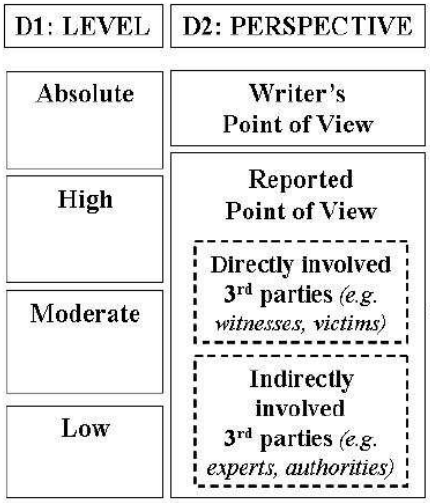

Figure 1: Dimensions 1 and 2, Certainty Categorization Model.

The Focus dimension is illustrated with sentences (4) and (5): sentence (4) is an example of an abstract information and sentence (5) is a factual information.

(4) *In Iraq, the first steps* ***must be taken*** *to put a hard-won new security council resolution on arms inspections into effect.*

(5) *The settlement* ***may not fully compensate*** *survivors for the delay in justice, …*

At last, the Time dimension is understandable without examples.

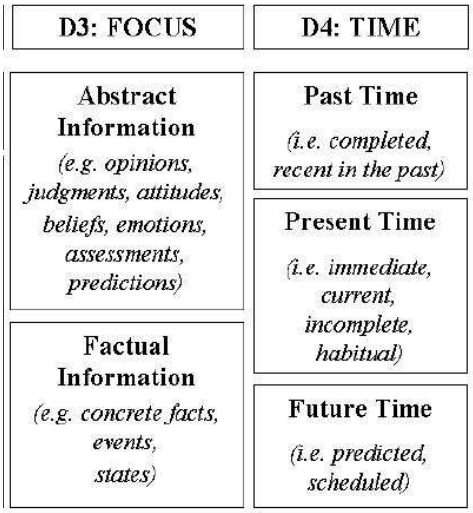

Figure 2: Dimensions 3 and 4, Certainty Categorization Model.

This model was developed for manual annotation. For our objective, the identification of the reported point of view is necessary. For example, if an event is reported by the government spokesperson or by the leader of the rebels, the user may associate quickly the uncertainty of the reported event, according to the reliability of the source. Also, we think that the reality of an event is lacking. For example, if we have "Laurent Gbagbo didn't go to Italy.", the sentence deals with the Move event between Laurent Gbagbo and Italy, but in a negative way. This negative expression has to be captured in order to keep this nuance for further treatment (expansion of a knowledge base). At last, we want to automatically extract those uncertainty and reality information.

4. Event Extraction with Uncertainty and Reality Detection

4.1 Enrichment of the Rubin et al. uncertainty model

We present here our model, which is an enrichment of the Rubin et al. uncertainty model. It includes the identification of the local source, which is necessary for an end-user to evaluate the reliability of the reported discourse. It also takes into account the reality or unreality of an information which is specified in the source text, rather than the Focus dimension. For example, if we have "Laurent Gbagbo didn't go to Italy.", the sentence deals with the Move event between Laurent Gbagbo and Italy, but in a negative way. But it's not an opinion, or an abstract information as in the Focus dimension. Here is the model of uncertainty and reality that we have defined in a context of textual information extraction.

D1 LEVEL: High, Moderate, Low.

D2 PERSPECTIVE: Writer's point of View, Reported Point of View.

D4 TIME: Past Time, Present Time, Future Time.

D5 REALITY: Assertion, Negative.

D6 SOURCE NAME. (only when D2= Reported Point of View)

Figure 3: Our Uncertainty and Reality Model.

In our model, we did not keep the Absolute value of the Level dimension, because in our current implementation we only define three levels of uncertainty. We will have to analyze whether this Absolute value is necessary or not. Also, we did not keep the distinction between the Directly involved 3rd parties and the Indirectly involved 3rd parties for the Perspective dimension, as it seems to be too hard to identify it automatically from text. Finally, we did not keep the Focus dimension, as it was not useful according to our needs.

4.2 Linguistic patterns for uncertainty and reality detection

We have developed linguistic patterns to identify the values of uncertainty in texts according to our five dimensions. This work was done for French, but could be realized for English also. Here is a subset of the linguistic knowledge used to identify uncertainty and reality.

Categories	Linguistic Forms	Dimensions and associated values
Adjectives	*douteux, incertain, peu probable*	Level: Low
	présumé, supposé,	Level: Moderate
	vraisemblable, probable, possible, envisageable, envisagé,	Level: High
Verbs	*dire, déclarer, annoncer* *penser, croire, douter, hésiter, …*	Level: Moderate
Expressions	*selon toute vraisemblance, sans doute, à ce qu'on dit,* *il se peut que, il paraît*	Level: Moderate
Structures	*selon, d'après , de source(s)* « … »	Perspective: Reported Point of View
	si (except when followed by an adverb)	Level: Moderate
	aller + infinitive	Time: Future Time
	ne, n' (except « *n'importe* »)	Reality: Negative

Table 1. Linguistic Knowledge used in patterns.

Those words or expressions are used in linguistic patterns in order to identify the sentence part concerned by the uncertainty. Our patterns are finite state automaton, defined with Intex, a linguistic development tool [11]. Here is an example of pattern, which annotates the moderate level of uncertainty.

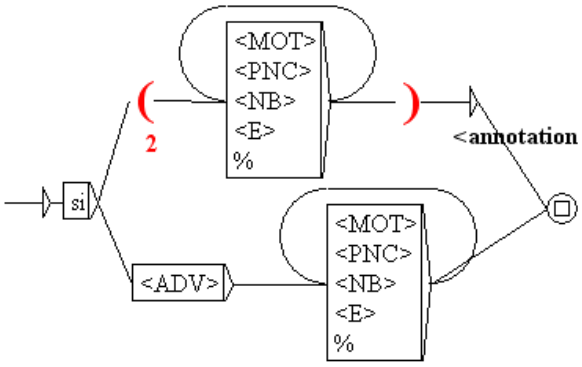

Figure 4. Linguistic Pattern Example.

In the first line of this pattern, "*si"* ("if" in French) introduces a part of sentence annotated with a Moderate level of uncertainty (we can see the beginning of the annotation expression at the right). In the second line no annotation is added when "*si"* is followed with an adverb (<ADV>). In this pattern, a sequence of any word is represented by the loop on boxes with <MOT>+<PNC>+<NB>+<E>+%.

4.3 Combination of Uncertainty and Event Extraction

Usually, uncertainty or other non-factual information are modeled by linguists but not used for automatic detection. On the other hands, some tools are dealing with the information extraction, such as relation or event extraction, but without capturing the nuances of the discourse. In our approach we have combined the automatic detection of events from text with the detection of the uncertainty and reality associated to the events, in order to keep all the contextual information. Our aim is to identify not only the uncertainty, but also the reality or unreality of the event, as sometimes discourses are dealing with events that never occur.

In practical terms, we first associate certainty characteristics to a part of sentence or a whole sentence of a text. Then, we apply event extraction linguistic patterns to extract events from text. When we identify an event, we add it all the certainty characteristics that have been associated to the sentence where the event was found.

5. Implementation

5.1 The WebContent Project Context

This work is done in the French ANR WebContent project context. The objective of WebContent is to provide a platform with services for text analysis for strategic watch applications. It is based on semantic web paradigm, so WebContent uses web services and Ontologies. Several web services (language identification, named entity identification, crawling, classification…) are developed by partners (CEA, EADS, Exalead, INRA, INRIA…) and are used in the strategic watch applications of the project (economic watch in aeronautics, strategic intelligence, microbiological and chemical food risk, watch on seismic events).

5.2 The Event Extraction Web Service

We have developed a web service for the event extraction, which implements also the detection of uncertainty. It is developed in Java. The web service analyses texts by applying linguistic patterns thanks to Intex. Some linguistic patterns are associated to uncertainty, some are associated to events. Linguistic patterns associated to events are built previously with the SemPlusEvent tool. This tool, which is the last version of the SemPlus tool, aims at easing the capture of event linguistic patterns from examples thanks to a learning algorithm [5]. At runtime, to configure the web service, we need an ontology of the domain with instances, which is transformed into dictionaries compatible with SemPlus. Then, for each input text, the web service takes in input a MediaUnit structure, which is a WebContent format containing texts, and adds semantic annotations in RDF to this MediaUnit which is provided as input. Here is an example to illustrate this implementation. We have analyzed the following sentence:

Selon des témoins, Laurent Gbagbo aurait rencontré Alassane Ouattara.

The analysis of this short text produces the following RDF annotations (figure 4, in thick, important information).

The first part of this RDF annotation contains all the characteristics of the Uncertainty#1, which is associated to the Event#0 described in the last RDF annotation. In this event, Personne56 is "*Laurent Gbagbo*", and Personne4 is "*Alassane Ouattara*". Those annotations are related to the corresponding textual segments via others RDF descriptions.

```
...
<rdf:RDF ...> <rdf:Description
rdf:about="weblab://InstanceCandidate//
Uncertainty#0">
<onto:Level>Moderate</onto:Level>
<onto:Time>Past time</onto:Time>
<onto:Reality>Assertion</onto:Reality>
<onto:Source>des témoins</onto:Source>
<onto:Perspective>Reported point of
View</onto:Perspective>
</rdf:Description> </rdf:RDF>
...
<rdf:Description
rdf:about="weblab://InstanceCandidate/Event#0">
<rdf:type rdf:resource="RENCONTRE"/>
<onto:Agent>http://www.owl-
ontologies.com/RCI.owl#Personne56</onto:Agent>
<onto:Patient>http://www.owl-
ontologies.com/RCI.owl#Personne4</onto:Patient>
<onto:related_uncertainty>weblab://InstanceCandidate/U
ncertainty#0</onto:related_uncertainty>
</rdf:Description>
...
```

Figure 5. RDF Annotations produced by the web service.

This web service is used in the economic watch in aeronautics and strategic intelligence applications developed in the WebContent project.

5.3 Evaluation

We have carried out a first evaluation of this work. It was done on a small corpus of 5 French articles dealing with news. This corpus contains 40 reported discourses, 25

uncertainty, 8 negation and 4 future. Here are two sentences from the corpus:

Selon lemonde.fr, les enregistreurs de vol de l'Airbus A330 qui s'est abîmé le 1er juin dans l'Atlantique avec 228 personnes à bord auraient été localisées par les navires de la marine française.[1]

Mir Hossein Moussavi et Mehdi Karoubi estiment que le vote a fait l'objet de vastes fraudes.[2]

Here are the results of the evaluation:

	Recall	*Precision*	*F-measure*
Perspective	0.73	0.95	0.83
Source	0.48	0.98	0.64
Level	0.84	0.58	0.69
Reality	0.88	1	0.94
Time	1	1	1

Table 2. Evaluation results.

Some reported discourses were not identified as the verbs used to introduce them were not taken into account (*commenter, voir,…*). Also, we work at the sentence level. Some sentences were not associated to reported discourses because when several sentences are in a reported discourse, only the first one, introduced with ", is identified. A similar situation occurs at sources identification: sometimes sources appear out of the sentence containing the reported discourse. In this case, they are not identified by our approach. Previously, we considered that a source could not contain ".", but we have to take into account sources such as web sites (lemonde.fr).

We considered at last that all reported discourses were associated to a "Moderate" certainty, according to the writer's point of view. But, we have observed that sometimes the quotation marks are also used to introduce a sentence pronounced by another person, without certainty or uncertainty.

6. Conclusion

We have presented a model that allows the precise characterization of uncertainty in a context of information extraction from texts. We have described our approach based on linguistic patterns and detailed a part of the linguistic knowledge for French analysis. This approach is used in a web service that produces RDF annotations containing the level of uncertainty, the time, the reality, the perspective and the source characteristics. Currently those annotations are used in strategic watch applications.

To improve our approach, we will need to take into account the results of the evaluation. We also will need to compare our annotations to the inputs and needs of fusion tools that may fused events according to their uncertainty characteristics.

7. Acknowledgements

This work was done in the WebContent ANR French Project context (http://www.webcontent.fr).

8. References

[1] ACE 2007, http://www.itl.nist.gov/iad/894.01/tests/ace/2007/.

[2] A. Auger, J. Roy. Expression of Uncertainty in Linguistic Data, in *Fusion 2008*, Cologne, Allemagne.

[3] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.

[4] FrameNet, http://framenet.icsi.berkeley.edu/.

[5] B. Goujon. Relation Extraction in an Intelligence Context, in *LangTech 2008*, Roma, Italy.

[6] M. Hecking. System ZENON – Semantic Analysis of Intelligence Reports, in *LangTech 2008*, Roma, Italy.

[7] M. Naughton, N. Kushmerick, and J.Carthy. Event Extraction from Heterogeneous News Sources, in *AAAI 2006 Workshop on Event Extraction and Synthesis*, Boston.

[8] T. Poibeau. Extraction d'information à base de connaissances hybrides, PhD, 2002.

[9] V. L. Rubin, E. D. Liddy, N. Kando. Certainty Identification in Texts: Categorization Model and Manual Tagging Results Computing Attitude and Affect in Text: Theory and Applications, The Information Retrieval Series, Springer Netherlands, vol. 20. pp. 61-76.

[10] R. Saurí, M. Verhagen, J. Pustejovsky. Annotating and Recognizing Event Modality in Text, *FLAIRS 2006*, Floride.

[11] M. Silberztein, 1999. INTEX: a Finite State Transducer toolbox, in *Theoretical Computer Science #231:1*, Elsevier Science.

[1] According to lemonde.fr, the recorders of the flight of the Airbus A330 which … should have been located by the ship of the French navy.

[2] Mir Hossein Moussavi and Mehdi Karoubi consider that the vote was the object of massive frauds.