Semantically-driven Competitive Intelligence Information Extraction: Linguistic Model and Applications

Iana Atanassova, Gan Jin, Ibrahim Soumana, Peter Greenfield and Sylviane Cardey

CRIT - Tesnière, Université de Bourgogne Franche-Comté
30 rue Mègevand, 25000 Besançon, France
Email: { surname.name } @univ-fcomte.fr

***Abstract*—In a competitive environment and in the current context of rapid technological advances, competitive intelligence is a key strategic need in the private sector and requires the development of Web content tools capable of robust and semantically-driven text classification. In this paper, we present a method for the information extraction and semantic classification of text segments. Our approach to text processing makes use of linguistic clues to populate an ontology of competitive intelligence. We have developed a method for the automatic identification and classification of sentences into predefined semantic classes by using linguistic models and a knowledge-based approach. This method has been tested on a dataset of journal articles in horology and aeronautics, and can be extended to other domains. The tool that we have developed is part of the WebSO+ platform for competitive intelligence. We present the overall methodology for annotation and information extraction, our experimental protocol and the results obtained from the evaluation.**

Keywords–Web content; Information Extraction; Competitive Intelligence; Sentence Classification; Semantic Annotation; Linguistic Model.

I. INTRODUCTION

Today, innovation in the private sector is conditioned by the context of a highly competitive international environment and rapid technological advances. For these reasons, competitive intelligence has become a key strategic need, i.e., companies have to monitor constantly the new technologies in their domains, but also market changes and emerging trends, the activities of competitors and partners, etc. [1]. In addition to this, the amount of information generated daily in each particular domain may lead to information overload. The traditional information sources have been enriched by new media such as social networks and customer opinions on the Web. For these reasons, competitive intelligence is becoming more and more costly in terms of the time and human effort to process the information. For example, [2] analyzes recent practices in European firms and arrives at the conclusion that competitive intelligence "has grown well beyond competitors to include customer related intelligence, technology, market, etc.".

Our work tackles the problem of the development of Web content tools capable of the automatic processing of news articles and user feedback in order to extract and classify specific types of information relevant to competitive intelligence. The major objective is to facilitate and accelerate the task of competitive intelligence in companies by providing tools for the efficient monitoring of both published information sources (e.g., news articles) and also customer feedback. We have developed a method for the automatic identification and classification of sentences into predefined semantic classes by using linguistic models and a knowledge-based approach.

As observed by [3], the majority of commercial applications in information extraction make use of rule-based approaches, while statistical and machine-learning (ML) approaches are widely used in the academic world. Table I gives the proportions of the use of various approaches employed in academic research and in the private sector (vendors of products for Information Extraction) for the implementations of entity extraction. These results were obtained by [3] in a study of conference papers over a 10 year period. Among the reasons for which the rule-based approaches dominate the commercial market is their advantage in terms of flexibility and traceability. The most recent efforts by the research community in this field are directed towards providing standardized rule languages and rule editing tools [4][5].

TABLE I. TYPES OF APPROACHES USED FOR INFORMATION EXTRACTION

	Rule-based (%)	*Hybid (%)*	ML based (%)
Scientific articles	3.5	21	75
Large vendors	67	17	17
All vendors	45	22	33

The problem of text classification in general has been the subject of many studies [6], most of which can be considered as document retrieval tasks in the sense that they work at the level of the document. In our approach, we focus on sentence extraction and sentence classification, which better correspond to the user need, i.e., to identify the exact information related to the competitive intelligence task, rather than retrieving the document that contains such information.

In this paper, we describe our approach to text processing which makes use of linguistic clues to populate an ontology of competitive intelligence. The module for Information extraction that results from this approach has been implemented as part of the WebSO+ platform for competitive intelligence that provides an environment for monitoring various types of information sources. The textual data is obtained by a separate module that scrapes lists of web sites and customer review databases on a daily basis. In this paper, we focus on the automatic classification module that is used for information extraction and allows filtering relevant text segments amongst

 ISBN: 978-1-61208-707-8

the large mass of data retrieved daily, and which are presented to the end user.

The experiment that we report here has been carried out on two specific industrial sectors which are horology and aeronautics. These sectors were chosen because they are well developed and strategically important in the Region of Franche-Comté in France and in Switzerland. The linguistic resources have been developed specifically for the processing of news articles and customer feedback in these two domains. However, our methodology, which relies on a general approach, can be applied to other domains after adapting the linguistic resources to take into consideration the new domains. These resources consist of linguistic rules that follow a grammar specifically designed for this purpose and which make use of lists of regular expressions.

The rest of the paper is organized as follows. In the next section, we present the linguistic model and the overall approach for automatic classification of text segments. In Section III, we give the details of the evaluation that has been carried out on the classification module, and the results. In Section IV, we discuss some of the difficulties in the processing and give examples of errors and possible solutions. Section V presents our conclusion and future work.

II. Linguistic Model and Automatic Classification Method

In this section, we first present the semantic subclasses that form our ontology of competitive intelligence. These subclasses are then used for the classification of sentences and information extraction based on our linguistic model.

A. Ontology of Competitive Intelligence

The objective of our method is firstly to identify sentences that are relevant to the task of competitive intelligence, i.e., that contain explicit information on competitors, market trends, innovations etc. At the same time, we classify these sentences into several subclasses, which are presented in Figure 1, where the names of the subclasses are given in English, and the French translation is in brackets.

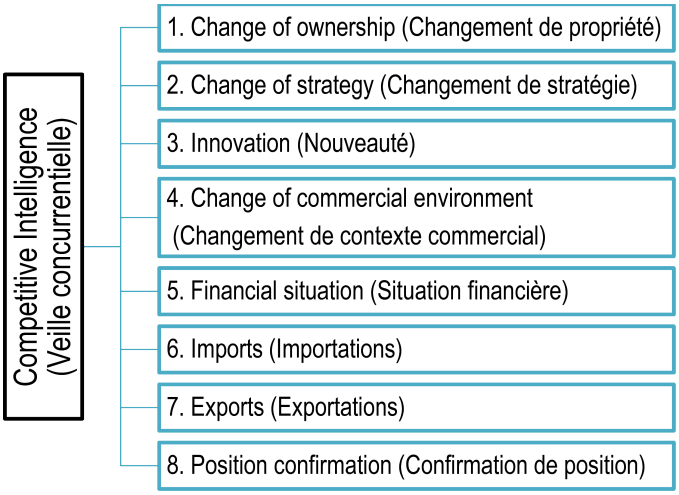

Figure 1. Ontology of Competitive Intelligence: subclasses used for the classification of sentences

Each of the subclasses can be expressed in texts using various expressions and linguistic structures. The table II presents examples of sentences that correspond to some of the subclasses. These examples have been extracted from articles that belong to the corpus described in the following section.

B. Linguistic Model

To develop the linguistic resources, we have analyzed the ways in which the subclasses are expressed in texts by studying a corpus of about 2,000 documents in French of different types: news articles, scientific publications, patents, customer feedback, etc. We then established sets of linguistic structures that are directly implementable and that allow identifying these types of information in new texts. This approach uses the SyGuLAC theory that stems from the microsystemic approach and discrete mathematics [7] in order to propose tools for linguistic analyses and their generalization.

The analysis is considered from the point of view of sense-mining in order to make possible the identification of relevant information in texts. Unlike data-mining systems that search for keywords in a sentence or in a text, in order to identify relevant information, we propose working at the level of sense which is present at all levels of analysis: lexical, syntactic and semantic and their intersections: morpho-syntax, lexico-syntactic-semantics, etc. [8]. However, our aim is not to construct a model that describes the language in its entirety with a global representation of its different levels separately: lexical, syntactic, mophological, semantic. Rather, we concentrate only on one specific objective at a time, which is, in the current study, the identification of information related to competitive intelligence. Thus, in our approach we consider only the elements that are necessary and constitutive of the problem at hand, which can be lexical, morphological, syntactic, etc. in nature. These elements are represented in linguistic structures that are directly implementable in terms of sets of regular expressions or other features that are identifiable in strings [9][10].

C. Implementation

The linguistic structures are represented in our systems as regular expressions following a grammar that was designed specifically for this task. As an example, the structure in Figure 2 represents a part of a subclass ”1. Change of ownership” of textual segments in French in the domain of horology. In this structure, several operators are used, e.g., *Verbe, opt*, that correspond to abstract representations in our model. The linguistic structure uses microsystems that are defined in order to tackle one specific problem. The specifications of the operators and constraints in the grammar are defined as in [7]. Several hundred such structures are associated with each of the above subclasses.

This architecture for information extraction has several advantages:

- the linguistic resources (structures) are independent of the processing model and the implementation of the information extraction engine;
- this methodology can be adapted and used in domains that correspond to precise needs in industry, where machine learning is impossible due to the lack of large scale corpora;

 ISBN: 978-1-61208-707-8

TABLE II. EXAMPLES OF SENTENCES THAT CORRESPOND TO THE SUBCLASSES OF COMPETITIVE INTELLIGENCE

Example	*Class*
Le groupe horloger hispano-suisse Festina a repris les actifs de la société neuchâteloise Technotime.	1.
Alors que les groupes de luxe n'ont eu de cesse ces dernières années de consolider - quand ce n'est pas posséder - leur réseau de distribution, ils comptent désormais sur des blogs ou autres sites spécialisés pour effectuer du e-commerce.	2.
Le suisse Longines présente une montre équipée d'un mouvement à pile nouvelle génération.	3.
Car aujourd'hui, c'est certain, l'amateur n'est plus un collectionneur affectionnant les mécanismes comme dans le passé, mais un adepte averti de la valeur des choses.	4.
L'année 2017 devrait bien se présenter pour Tag Heuer, a noté Jean-Claude Biver, se disant toutefois prudent face aux incertitudes économiques et géopolitiques du monde et tablant sur une croissance "à un chiffre" pour la marque.	5.
Les exportations horlogères suisses ont continué de reculer en février, accusant une baisse de 10% à 1,5 milliard de francs suisses (1,3 milliard d'euros), a annoncé mardi la Fédération de l'industrie horlogère suisse (FH).	7.
"Nous avons consolidé notre quatrième position dans l'horlogerie suisse en matière de chiffre d'affaires, derrière Rolex, Omega et Cartier", souligne celui qui entré chez Longines en 1969 et qui dirige la société depuis 1988.	8.

listEH/opt(Arg1) + Verbe(v-ach) + opt((listEH)/Arg2) + (...) + opt(Ctxt)

Figure 2. Example of a linguistic structure for the subclass "1. Change of ownership"

- the linguistic analysis is based essentially on the structures that are defined by linguists, unlike in other approaches that rely on machine learning of keyword distributions;
- there is a complete traceability of all the linguistic structures involved in some task resolution, which means that the sources of errors can be identified and corrected by modifying the erring structures;
- the modification or the improvement of one linguistic structure can be done independently of the other linguistic structures, which means that the performance of the system can be incremented fairly easily by correcting existing structures to improve the precision or adding new structures to achieve higher recall.

The linguistic analyses take into consideration large context spans in the textual segments. In fact, from a linguistic point of view, we know that the presence of words or particles very far away in the linear representation of a sentence can have a significant impact on the overall meaning. For this reason, the linguistic structures that we use take into consideration the entire sentences. This approach to language modeling presents a considerable advantage for the description of linguistic phenomena compared to other methods that are inspired by the 'bag of words' model.

III. EVALUATION

We have performed an evaluation in order to quantify the capacity of our system to correctly identify and classify sentences that contain information relevant to the task of competitive intelligence for French. In this section, we describe the method that we adopted for this evaluation and the results obtained.

A. Corpus

We have constructed a corpus of news articles in the two sectors of horology and aeronautics in French. The corpus was collected manually by exploring online journals and by using search engines to identify relevant sources and articles on the Web. The sources of the articles are of two types: general newspapers and sites (e.g., Le Monde, Figaro, Le Parisien, Le Point, Les Echos, Le Temps (CH), Tribune de Genève), and specialized magazines and sites in the domains of horology, aviation, new technologies, stock exchange (e.g., www.montres-de-luxe.com, fr.worldtempus.com, www.meltystyle.fr, www.journal-aviation.com, last access 1/3/2019) that are published both in France and in Switzerland at the beginning of 2017. The evaluation corpus contains a total of 45 documents and 1,027 sentences. Table III gives more details of the size of the corpus.

TABLE III. DESCRIPTION OF THE EVALUATION CORPUS

Domain	*Documents*	*Sentences*
Horology	30	745
Aeronautics	15	282
Total	45	1 027

The types of documents and their format in the evaluation corpus are similar to the real-case use for which we have designed the information extraction engine.

B. Evaluation Protocol

We compared the results of the automatic classification with a gold standard obtained by manually classifying the sentences in the corpus. The evaluation was done following the four stages described below.

Stage 1. Automatic segmentation of the corpus into sentences.

Stage 2. Manual classification of all sentences. This was done by students and researchers in the domain of Natural Language Processing. Each sentence in the evaluation corpus was examined and identified as relevant or irrelevant to the task of competitive intelligence. Then, all relevant sentences were assigned one of the 8 subclasses.

Stage 3. Automatic classification of the sentences in the corpus by running our classification module.

Stage 4. Calculation of the precision (P), recall (R) and F-measure (F) [11].

C. Results

Figure 3 shows the number of sentences in the evaluation corpus that were manually classified into each subclass considering the two domains of horology and aeronautics. We observe that the subclass "3. Innovation" contains most of the sentences related to horology, and the subclasses "3. Innovation" and "5. Financial situation" contain the largest sets of sentences in aeronautics.

 ISBN: 978-1-61208-707-8

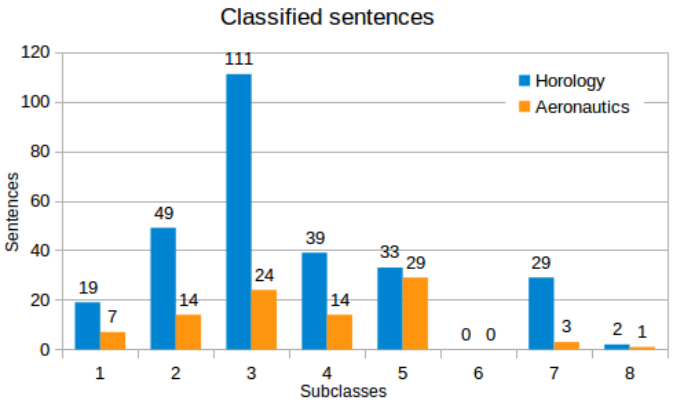

Figure 3. Distribution of the sentences according to the 8 subclasses in the manual classification

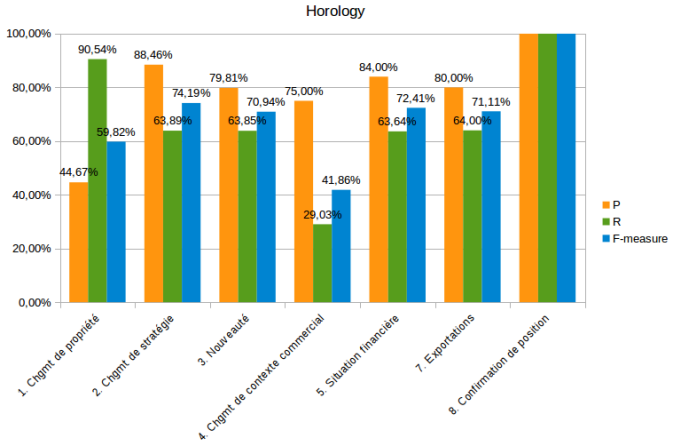

Figure 4. Results of the evaluation in the domain of horology

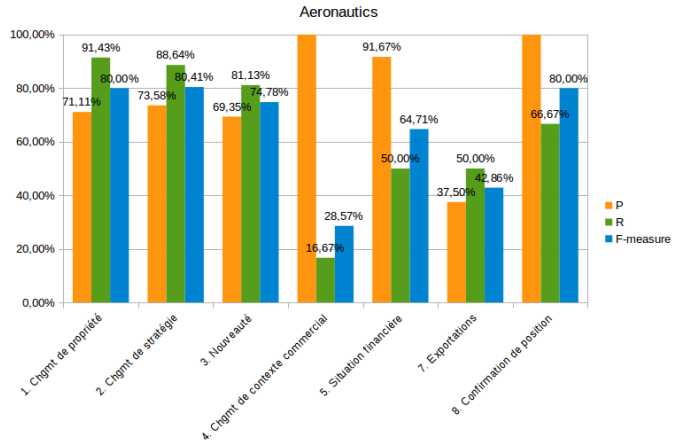

Figure 5. Results of the evaluation in the domains of aeronautics

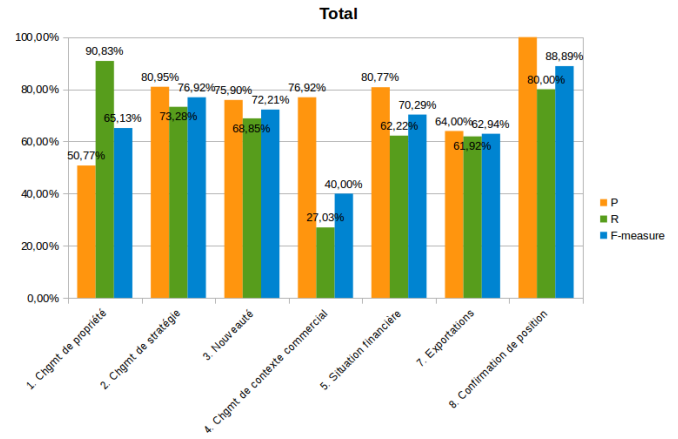

Figure 6. Results of the evaluation in both domains of horology and aeronautics

Figures 4 and 5 present the values of the precision, recall and F-measure calculated by comparing the manually and automatically classified sentences.

In Figure 4, we observe that in horology the values for the precision are high (above 75%) for all subclasses except for ”1. Change of ownership”. This score is due to the fact that this first subclass is complex as it contains different types of sentences that express various kinds of merger-acquisition transactions, as well as liquidations and joint ventures. To improve the system’s performance for this subclass, the corresponding linguistic structures can be identified and corrected. At the same time, the value of the recall for this subclass is above 90%. Considering the subclass ”4. Change of commercial environment”, the values of the recall are low (around 29%) because these kinds of changes can be expressed in many various ways. A larger number of linguistic structures should be considered to improve the coverage.

In Figure 5, the values of the precision are also quite high except for the subclass ”7. Exportations”. In fact, the manual classifications for this subclass show that in some cases a confusion can be made between the subclasses ”7. Exportations” and ”1. Change of ownership” for sentences that express large investments or purchases in aircraft companies. The value of the recall for the subclass ”4. Change of commercial environment” is low and, as in the domain of horology, this subclass needs a more comprehensive set of linguistic structures.

Figure 6 presents the results of the evaluation for both the domains of horology and aeronautics. The overall performance of the system is satisfactory as the majority of the precision and recall values are above 70%. As we have noted, the subclasses of ”1. Change of ownership” and ”4. Change of commercial environment” still need some improvement.

IV. Discussion

The results presented above show that the methodology that we used based on sets of linguistic structures is adequate for the identification and the classification for most of the subclasses. However, improvements can be made for 2 subclasses. In this section, we consider some of the typical errors and discuss the possible improvements that can be made using our methodology.

Table IV gives some typical examples of errors that were extracted from the evaluation corpus. We have analyzed all sources of errors and listed the causes and actions needed

 ISBN: 978-1-61208-707-8

TABLE IV. EXAMPLES OF SENTENCES THAT WERE NOT CORRECTLY IDENTIFIED OR CLASSIFIED BY THE SYSTEM

Example	*Automatic class*
Le siège est installé à La Chaux-de-Fonds, dans le canton de Neuchâtel, à quelques dizaines de kilomètres de la frontière du Doubs ; elle compte trois usines dans le Jura suisse et un centre de recherche à Palo-Alto, en Californie.	1.
Le rythme ultra-cadencé des renouvellements de produits, de caducité des composants et la chute des prix ont ainsi causé la faillite de Pebble, pionnier du secteur.	not identified
L'ascension rapide et le potentiel d' Akrone ont séduit Christophe Courtin, qui, avec son fonds Courtin Investment, vient de prendre 25 de son capital.	not identified
Cela explique, comme le souligne Alain Zimmermann, CEO de Baume & Mercier, pourquoi toutes les marques de luxe recentrent depuis peu leur intérêt sur des modèles à forte identité.	not identified
Et de rappeler que lorsque Tag Heuer a lancé sa montre connectée en 2015, 4 millions d'impressions Internet ont été enregistrées en deux jours.	4.
Ali Nouri – c'est son nom – mise beaucoup sur cette clientèle potentielle et il ne lancera la production de ses premiers modèles, en Chine, une fois seulement qu'il aura engrangé suffisamment de commandes.	3.
Doté d'un calibre Dior Inversé et d'une masse oscillante fonctionnelle visible à l'avant du cadran, le garde-temps évoque par sa structure mécanique les tournoiements d'une délicate robe de grand soir.	not identified
D'autant que les acheteurs, en particulier masculins, se projettent aisément et font de leur montre une sorte de talisman qui les transforme en héros d'une saga qu'ils écrivent dans leur tête.	not identified
Une relance constatée aussi par Swatch Group, dont un tiers de l'activité se fait en Chine.	not identified
L'ascension rapide et le potentiel d' Akrone ont séduit Christophe Courtin, qui, avec son fonds Courtin Investment, vient de prendre 25 de son capital.	not identified

to improve the performance of the system. These can be summarized in the following several points:

- Noise in the automatic classification due to some structures that are "too general" and identify a large number of sentences. These structures can be improved in order to identify only relevant sentences by adding new lists and constraints.
- Silence in the system due to the presence of rare words, expressions or neologisms in some sentences. This problem can be solved by adding new linguistic structures.
- Some sentences make use of figurative language (metaphors, comparisons, etc.) in the news articles. This problem is difficult to tackle in general, but the most frequent cases can be studied and the linguistic structures can be adjusted to take into consideration some of these phenomena.
- Errors related to the sentence segmentation: the limits of the identified segments are of crucial importance for the application of the linguistic structures as they take into consideration entire sentences. In some rare cases the segmentation is not correct and this can be improved.
- Errors due to the use of negation in the sentences: negation in French is expressed in most cases by two words that surround the verb form (*ne ... pas, ne ... aucun, ne ... jamais, ...*). Such cases need more constraints.

Our approach has been developed to respond to the specific needs of competitive intelligence in the private sector, and the choice of the classes addresses these needs. While other methods exist for sentence classification, such as machine learning or neural network approaches using pretrained word embeddings [12], such methods depend heavily on the availability of large annotated datasets that are necessary for the training of the model. Obtaining such datasets with good quality annotations is expensive. To our knowledge, no such datasets exist in the field of competitive intelligence, and therefore the use of the latest deep learning approaches in this contexts is not practically applicable. In the approach that we propose, the major effort is concentrated on the development of the linguistic model for the classifier rather than the manual annotation of a dataset. If the results are satisfactory, this method could be used as a bootstrap process to produce large annotated text corpora that could in turn be used as training datasets for neural network models.

V. Conclusion and Future Work

In this paper, we have presented an overall approach for the automatic classification of sentences based on sets of linguistic structures and its implementation for the task of competitive intelligence. We report on the results of the experimentation on news articles in two specific domains that are horology and aeronautics in French. The same methodology can be adapted to other domains if necessary.

Our classification module is part of a comprehensive platform for competitive intelligence, where the automatic classification helps users rapidly identify relevant information and thus deal with large volumes of data. In a real case scenario, hundreds of sources need to be scanned daily by the user. The capability of the system to highlight automatically classified sentences related to competitive intelligence plays an important role in diminishing the workload and enabling the user to digest ever larger amounts of information.

Our future efforts will be directed in two directions. Firstly, we aim to develop Semantic Web APIs, in order to render the data available through SPARQL requests and build new interfaces. Secondly, this methodology should be applied and evaluated on datasets of articles in other domains, such as the automobile industry and smart cities.

Acknowledgments

Part of this research has been funded by the FEDER (Fonds européen de développement régional) and selected by the French-Swiss programme Interreg V: WebSO+ project.

The authors thank Laurence Gaida and Philippe Payen de la Garanderie, members of the CRIT laboratory of the University of Bourgogne Franche-Comte, for their participation in the evaluation of the linguistic model.

References

[1] C. A. Bulley, K. F. Baku, and M. M. Allan, "Competitive intelligence information: A key business success factor," Journal of Management and Sustainability, vol. 4, no. 2, 2014, p. 82.

[2] J. Calof, R. Arcos, and N. Sewdass, "Competitive intelligence practices of european firms," Technology Analysis & Strategic Management, vol. 0, no. 0, 2017, pp. 1–14.

[3] L. Chiticariu, Y. Li, and F. R. Reiss, “Rule-based information extraction is dead! long live rule-based information extraction systems!” in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 827–832.

[4] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, “UIMA Ruta: Rapid development of rule-based information extraction applications,” Natural Language Engineering, vol. 22, no. 1, 2016, pp. 1–40.

[5] W. Wang and K. Stewart, “Spatiotemporal and semantic information extraction from web news reports about natural hazards,” Computers, environment and urban systems, vol. 50, 2015, pp. 30–40.

[6] M. Allahyari et al., “A brief survey of text mining: Classification, clustering and extraction techniques,” arXiv preprint arXiv:1707.02919, 2017.

[7] S. Cardey, Modelling language, ser. Natural Language Processing Series. John Benjamins Publishing Company, 2013.

[8] S. Cardey et al., “A model for a reliable automatic translation, the TACT multilingual system, LISE project (Linguistics and Security),” in Proceedings of WISG’09, Workshop Interdisciplinaire sur la Sécurité Globale, Troyes, France, 2009.

[9] G. Jin, “A system for French-Chinese automatic translation in the domain of global security,” Ph.D. dissertation, University of Franche-Comté, Besançon, France, 2015.

[10] G. Jin, I. Atanassova, I. Souamana, and S. Cardey, “A model for multilingual opinion and sentiment mining,” in Conference TOTh 2017, Terminology & Ontology : Theories and applications, Chambery, France, 2017, pp. 283–287.

[11] C. J. Van Rijsbergen, “Foundation of evaluation,” Journal of Documentation, vol. 30, no. 4, 1974, pp. 365–373.

[12] Y. Zhang, S. Roller, and B. Wallace, “MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification,” arXiv preprint arXiv:1603.00968, 2016.

 ISBN: 978-1-61208-707-8