AUTOMATIC EXTRACTION OF SPATIO-TEMPORAL INFORMATION FROM ARABIC TEXT DOCUMENTS

Abdelkoui Feriel [1] and Kholladi Mohamed Khireddine [2]

[1]Department of Computer Science, MENTOURI 2 University, Constantine, Algeria
[2]HAMMA lakhdar , El oued University , Algeria

ABSTRACT

Unstructured Arabic text documents are an important source of geographical and temporal information. The possibility of automatically tracking spatio-temporal information, capturing changes relating to events from text documents, is a new challenge in the fields of geographic information retrieval (GIR), temporal information retrieval (TIR) and natural language processing (NLP). There was a lot of work on the extraction of information in other languages that use Latin alphabet, such as English,, French, or Spanish, by against the Arabic language is still not well supported in GIR and TIR and it needs to conduct more researches. In this paper, we present an approach that support automated exploration and extraction of spatio-temporal information from Arabic text documents in order to capture and model such information before it can be utilized in search and exploration tasks. The system has been successfully tested on 50 documents that include a mixture of types of Spatial/temporal information. The result achieved 91.01% of recall and of 80% precision. This illustrates that our approach is effective and its performance is satisfactory.

KEYWORDS

Arabic NLP, Information extraction, temporal data, spatial data, gazetteers, Gis.

1. INTRODUCTION

Due to the increasing number of Arabic content on the Web, an application is needed to exploit the large amount of information. In recent years, extracting and exploiting spatial and temporal information from text have been paid much attention in the fields of GIR and TIR and a lot of works have been done in mostly languages using Latin scripts, and have yielded satisfactory performances. But there were only little approaches that combine techniques, models, applications from those two fields in order to manage information with spatial characteristics that changes over time, or in other words, Spatio-temporal Information.

In addition 7to traditional IR capabilities supported by today's search engines, more and more search and exploration tools have emerged that focus on detecting and exploiting different types of so-called named entities in text documents. Named Entity Recognition (NER) is a technique of NLP which classify defined named entities such as organizations, persons, time and locations. Consequently, the need for techniques to automatically extract those named entities from unstructured text is increasingly important.

Building a system to extract Arabic information is a difficult task. Arabic language is a semitic language, it is well known for its complex morphology. In addition, Arabic does not have capital letters. Inversely, in the English language which allows mixed letter cases; some named entities

can be distinguished because they are capitalized. These include persons names, locations and organization [1].

There are a variety of tools to extract terms in languages such as French, English; some of them can easily be adapted with some minor modifications for the Arabic language. GATE [2] for example may be used in Arabic named entities extraction. EXIT [27] can be used to extract collocations but it needs a special Arabic tagger. Currently we are working on adapting GATE for our purpose. GATE is a language engineering environment developed at the University of Sheffield and has been used extensively for teaching and research since its first release in 1996. There is a set of reusable processing resources provided with GATE, which forms an information system named ANNIE (A Nearly- New IE system) [28]. ANNIE consists of the main processing resources for information extraction such as: tokeniser, sentence splitter, POS tagger, gazetteer, finite state transducer and orthomatcher. Another important component is the JAPE language (Java Annotation Pattern Engine) [29] which consists of a set of phases, each of which consists of a set of pattern/action rules.

Thus, in this paper we propose an approach for automated exploration and extraction of spatio-temporal information from Arabic natural language texts documents. We limit the geographical scope for Algerian territory.

The remaining of the paper is organized as follows. In section 2, we review the related works on geographical, temporal and geo-temporal search. In section 3, we present the system architecture and some of its components. In section 4 we present an example of a heuristic used to identify and combine the spatial and the temporal information and in section 5 we evaluate our system and compare its performances with the other related works.

2. RELATED WORK

The fast growing volume of spatial and temporal data pose fundamental GIScience challenges, ranging from conceptualization, representation, computation, and visualization. Associations between different spatial and temporal data sources and documents imply the development of novel retrieval mechanisms. In this section, we present some works related to our proposed work. In GIR, a key objective is to detect and capture location-based information from natural language text. Many studies on extracting geographic information from text documents have been proposed, and applied throughout previous years. [3, 4, 5, 6].

In the field of TIR, Research on temporal entity extraction in those languages that use Latin alphabet, such as English, German, French, or Spanish, uses local grammars, finite state automata [7,8,9], and neural networks [10] to detect temporal entities. These techniques do not work well for Arabic because of to the morphology and high ambiguity rate of Arabic. However, there are few approaches that consider both techniques of the combination and extraction of temporal and spatial information, some works done by [11] Focus on RSS feeds and extracts temporal and geographic information from such feeds. The work in [12] presented an approach that combines temporal and geographic information extracted from documents and recorded in temporal and geographic document profiles. [13] Presented a method for capturing the spatio-temporal patterns of hazard-related events from texts and to track the different kinds of events relating to both environmental and human perspectives over space–time. A hazard-based ontology has been presented to assist the spatio-temporal and semantic information extraction and retrieval process.
Instead other languages, the Arabic language is still not well supported in GIR and TIR, Arabic NER researches try continuously to develop and improve named entities recognition in the Arabic language, some efforts in [14] presents an SVM-based approach for Arabic NER with language generic and language specific features, resulting in a 10-30 point increase in F1 score over

baseline for person, location and organization named entity categories. Authors in [15] presented a system called Named Entity Recognition Arabic (NERA). The purpose of this system is to improve the rules based on named entity recognized by means of applying machine learning. Some works on [16, 17, 18] performed a rule-based approach for the extraction of explicit and implicit functional relations between person names and organizations for Arabic named entities. [19] Introduced a method for extracting Named Entities (NE) of locations and drugs entities and the relation among those two entities from Egyptian Arabic newswire.

Industry tools that extract temporal entities from Arabic texts exist [20, 21, 22].

3. THE PROPOSED APPROACH

Our approach combined many systems and corpora (Figure 1) , and a rule based approach is adopted to identify, extract and combine spatial and temporal information from Arabic text documents before such information is further utilized in search and exploration tasks, The proposed approach builds a model for automatically exploring patterns which indicates all type of spatial and temporal occurrences, for that, We adapt The General Architecture for Text Engineering (GATE) [23].

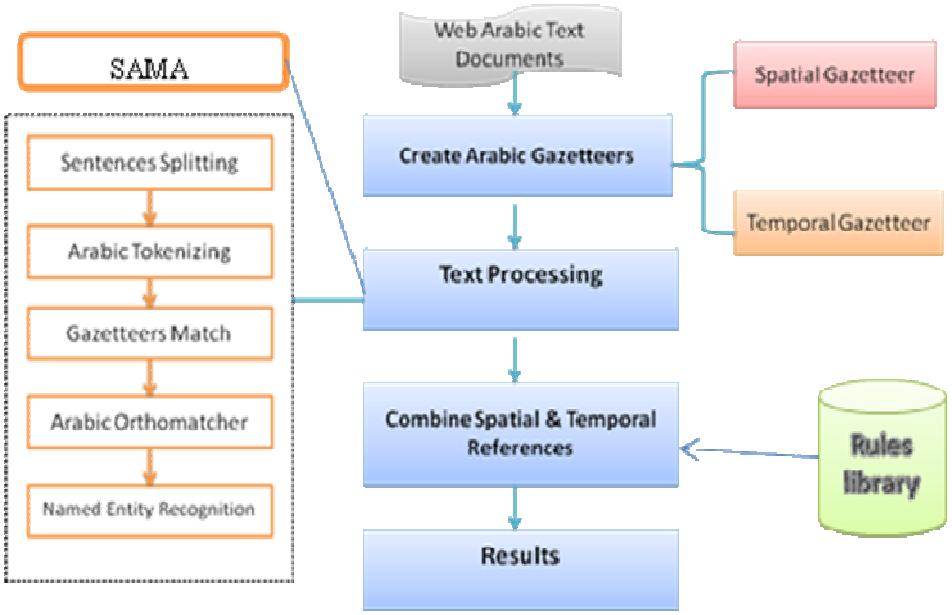

Figure 1. A framework for automatically extracting spatio-temporal information from Arabic text documents

Our approach consists of three phases, as shown in Figure 1. In the first phase," create Arabic gazetteers" information is collected, and the Arabic spatial and temporal gazetteers are created for contributing to the text matching steps, In the second phase, "text processing", different modules are applied to the text in order to recognize named entities. In the third phase "Extraction and combination phase", uses the rule library which consists of a set of Algorithms which are developed and implemented in Java. We describe those phases in depth, in the following.

3.1. Create Arabic Gazetteers

A gazetteer lists are plain text files with one entry per line, each list represents a set of names such as locations, names of cities…etc. This type of gazetteer is built manually.

3.1.1. Spatial Gazetteer

Place is a key concept in everyday life and reflects the way humans understand, experience and perceive their environment, Existing spatial databases and GIS are mature in representing space, but limited in representing place. Gazetteers are dictionaries of geo-referenced place names, and play an important role in GIR. Gate's default Arabic gazetteer doesn't cover Algerian geographic information; therefore we have created a spatial gazetteer of Algerian administrative division which consists of 3 levels: Province-level, dayrat-level (sub-prefectures) and commune-level. However, the problem we encountered during the implementation was the lack of reliable sources of Arabic data. At the first step, we have gathered public records from organizations such as Wikipedia, the Algerian postal company, and some local government institutions. From these sources, our gazetteer was initially populated with approximately 141,285 Algerian toponyms including names of cities, towns, villages, etc.

3.1.2. Temporal Gazetteer

In Arabic text documents, extraction of temporal entities is a hard task due to the morphology of the Arabic language and the way that the temporal entities are expressed. Human-written text is not consistent about the specificity of date expressions, some entities represent absolute time and dates such as 2007/07/20 or 2010 اوت 05 (05 august 2010), some entities represent relative time such as: بعد ستة أيام (after six days).

Extracting temporal entities with Gate's default gazetteer may limit the possible amount of information to extract. For that we have created an Arabic temporal gazetteer to complement GATE's default gazetteer, this provides temporal processing of 350 additional references, such as الصباح الباكر (early morning) or في ليلة الاثنين (on a Monday night) to extend the temporal annotation capabilities. Also after using the gazetteer we met some anomalies, among them the following examples in Table 1.

Table 1. Anomalies.

Cases	**Examples**
Some numbers with 04 are annotated as a year.	رقم: 3589 A number : 3589
only the suffix is annotated	2015 23 جوان June 23 2015.
wrong in the level of the date	2013 جانفي 75 75january2013
partially annotated	2012 افريل 30. 30 april 2012

We found the solution through using JAPE rules, the rule recovered separately the different components of the date (year, month and day), and restore the full date in a standard format than add it to the annotation type "Date". As it shown in Table 2:

Table 2. Date before and after normalization.

Date	**Normalized date**
1988-12-29	1988-12-29
2010	2010-12-31/01-01
2014April 28	2014 -04-28
2015 July	2015-07-31/01

3.2. Text Processing

Our system requires going through the pre-processing task before going to the analysis phase. The pre-processing resources of the Arabic language that we used are:

3.2.1. Sentence splitter

As the name suggests, it splits the text into its component sentences, and identifies sentence's boundaries.

3.2.2. Context

The context feature (the two previous and subsequent tokens) helps to determinate of extending or terminating named entities.

3.2.3. Arabic tokenizer

It divides the Arabic text into simple tokens such as words, punctuation and numbers...etc For example:

" تحتل الصحراء النصف الجنوبي من الجزائر"

"Desert occupies the southern half of Algeria"
will be tokenized as:

"<تحتل>, <الصحراء >, <النصف >, <الجنوبي>, <الغربي>, <من>, <للجزائر >, <.>"

3.2.4. Morphological analyzer

The Morphological analyzer Helps to group together words which express similar notions. The role of the Morphological analyzer in the Arabic language is to identify the morphemes of a word (Stem): the affixes (prefix, infix, and suffix) and the root. [24] A stem can be a noun, particle or verb. It can be composed of: One part (a root, for example: (ش ه ر)); Two parts (a root + a pattern, for example: (ه ر ش): root (ش ه ر) + a pattern; Three parts (a root + a pattern + affixes, for example: (الشهرين): root (ش ه ر) + a pattern + affixes (prefix (al) (ال), and the suffix (yan) (ين)). In our work, we used the Arabic Morphological Analyzer (the LDC Standard Arabic Morphological Analyzer (SAMA)), as we have mentioned previously only for temporal entities extraction; SAMA is a simple Arabic morphological analyzer which uses a rule-based system, it considers each Arabic word token in all possible prefix-stem-suffix segmentations, and lists all known possible annotation solutions, with assignment of all diacritic marks, morpheme boundaries

(separating clitics and inflectional morphemes from stems), and all Part-of-Speech (POS) labels and glosses for each morpheme segment [25].

3.2.5. Arabic main grammar

It allows using files containing the various rules and algorithms.

3.2.6. Arabic orthomatcher

For solving the problem of co reference.

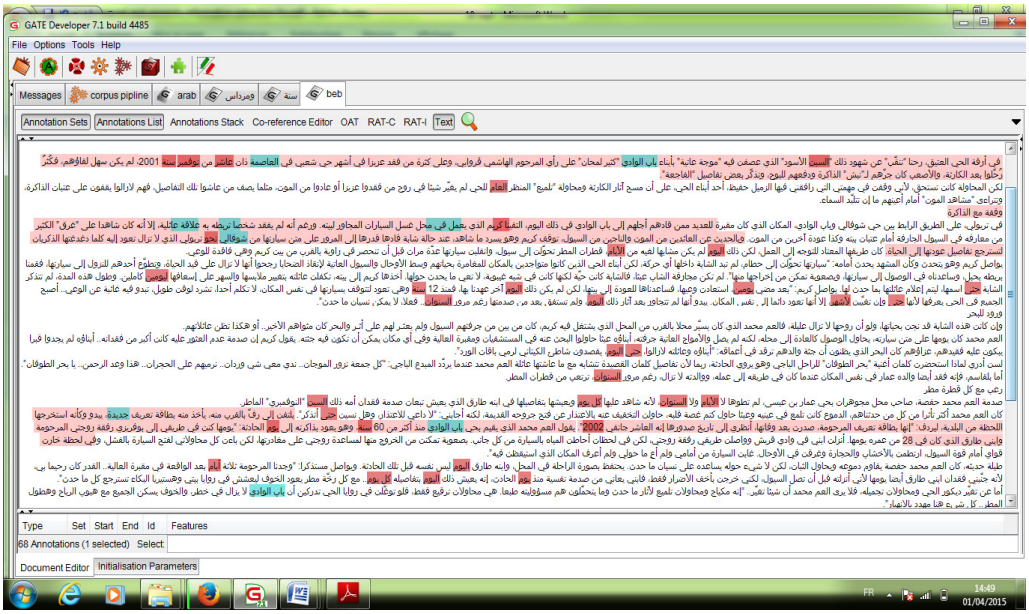

Figure 2. spatial and temporal annotation

3.3. Extraction and Combination

To extract and combine spatial and temporal information, a model is needed that precisely defines such information in documents (or rather the corresponding textual expressions) and how to combine them. Our model is based on the nature of textual phrases and how the spatial and temporal terms are presented in those phrases. For that, 04 cases are deducted in Table 3.

Table 3. Extraction of spatial and temporal information.

Cases	Examples
one spatial term and one temporal reference;	سنة كاملة مرت على فيضانات باب الوادي . A full year passed after floods of bab-el-oued

One spatial term and multiple temporal references;	اجريت المباراة في ملعب شاكر على الساعة 2 والدقيقة العشرون The match was held in Shaker Stadium at 2 and twenty minutes.
multiple spatial terms and a one temporal reference;	أكدت مصالح الحماية المدنية أنه لم تسجل اية خسائر مادية او بشرية جراء الزلزال الذي ضرب السبت جنوب شرق زموري بولاية بومرداس The interests of the Civil Protection confirmed that there have been no material losses or injuries from the quake, which struck Saturday southeast Zemmouri Boumerdes
Multiple spatial and multiple temporal references.	اندلعت التورة الجزائرية يوم 01/نوفمبر 1945 على الساعة 00.00سا.في جبال الاوراس ب باتنة Algerian revolution which was happened on 01 November 1954 at 00:00h on El-Aouress mountain in Batna town.

3.3.1. Algorithm

The spatio-temporal extraction process is dependent on the previous pre-prepared resources; the algorithm used in the extraction phase is as follows:

***Input**: document D, sentence E, spatial term S, temporal term T*
***Output:** combine S, T;*
Begin:
*Parse D, Read words **w** from the text*
***For** each sentence E in D do*
If** one **S** and one **T** in E **then
Combine (S, T)
If** one **S** and multiple **T** in E **then
Combine (S, T1), Combine (S, T2), Combine (S, T3)…
If** multiple **S** and one **T** in E **then
Combine (S1, T), Combine (S2, T), Combine (S3, T)…
***If** multiple **S** and multiple **T** in E*
then
*Check the left and right context of **S***
***If** there is a comma then*
Affect S1 to T, Combine (S1,T)
***Else** jump to S2...*
End

4. SYSTEM EVALUATION

This section describes the experiments conducted to confirm the effectiveness of our system. As preliminary experiment we chose newspapers texts. As our evaluation corpora, we have taken a set of around 70 news articles extracted from the Al-chorouk الشروق and al-khabar الخبر television Website [30, 31], and comparing the output against a manually tagged version of the text.

In order to evaluate the results, we employed recall and precision measures as our evaluation metrics. Detection precision refers to the fraction of the spatio-temporal entities correctly detected against the total number of spatio-temporal references that the system attempts to resolve . Detection recall refers to the fraction of the spatio-temporal entities correctly detected against the total number of all spatio-temporal references. The table bellow show the results obtained.

Table 4. Manual VS Automatic Annotation

All spatio-temporal references = 123	**Manual**	**Auto**
correct	105	99
incorrect	06	10
missed	08	03

From Table 5, we can see that for all the 123 spatio-temporal references, the results obtained by the human manually version are: 105 correct references, 06 incorrect references, and 08 missed references, against 99 correct references 10 incorrect references, and 03 missed references performed by the system, based on those results, we calculate the recall and the precision, as shown in Table 6.

Table 5. precisions of the 04 cases.

Cases	**precision**
one spatial / one temporal	**0.94**
One spatial / multiple temporal;	**0.89**
Multiple spatial /one temporal;	**0.79**
multiple spatial / multiple temporal.	**0.8**

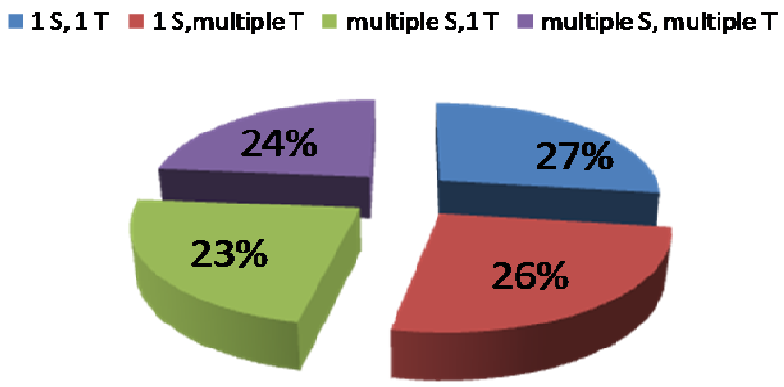

Figure 3. Precisions rates for each of the 04 cases.

Table 6. Comparison between the results of the presented system and other systems

Systems	**Precision**	**Recall**
Our system	0.80	0. 91
Wei wang's system [13]	0.86	0.88
David O'Steen's system [14]	0.84	0.77

From this comparison, it can be deducted that our system competes with the state of the art systems in terms of precision and recall.

4. CONCLUSION

In this paper, we presented an approach to automatically extract spatio-temporal information from Arabic text documents using NLP, GIR and TIR techniques. A set of steps was used to develop our system, starting from the creation of Arabic spatial and temporal gazetteers, to the text processing. At this step, this approach uses tow main components: the Arabic morphological analyzer SAMA, and the rule library which consists of set of grammatical rules. We made some experiments that show the possibility of obtaining the expected information in the returned results when using our approach. We have obtained as performance. 0.91% Recall, and 0.80% of precision, comparing with other related works, we can say that our approach is efficient and its performance is satisfactory.

Our future work will focus on the improvement of the rule library, gazetteers, for example including semantics by integrating ontologies, or spatial and temporal relations to treat more complex expressions.

REFERENCES

[1] Omnia. Z, et al, (2008)' A Novel Approach for Detecting Arabic Persons'", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.

[2] Maynard D, Cunningham H., et al , " A Survey of Uses of GATE" , Technical Report CS-00-06, Department of Computer Science, University of Sheffield, 2000.

[3] Mani, I., Anderson, D. and Hitzeman, J. (2006) A framework for interring spatial locations and relationships from text. National Center for Geographic Information & Analysis (NCGIA) Digital Gazetteer Research and Practice Workshop, http://ncgia.ucsb.edu/projects/nga/docs/mani-paper.pdf

[4] Jones, C.B. and Purves, R. (2008) Geographical information retrieval. International Journal of Geographical Information Science, 22(3): 219-228.

[5] Janowicz, K., Scheider, S., Pehle, T., and Hart, G. (2012) Geospatial semantics and linked spatiotemporal data-past, present, and future. Semantic Web, 3(4): 321-332.

[6] Machado, I. M. R., Alencar, R. O. D., Campos, R. D. O., and Clodoveu, A., D. (2011) An ontological gazetteer and its application for place name disambiguation in text. Journal of the Brazilian Computer Science, 17(4): 267-279.

[7] Li, H., Hu, Y., Gao, G., Shnitko, Y., Meyerzon, D., Mowatt, David: Techniques for extracting authorship dates of documents (December 2009).

[8] Koen, D.B., Bender, W: Time frames: temporal augmentation of the news. IBM Systems journal Journal 39 (July 2000) 597–61.

[9] Llid´o, D., Berlanga, R., Aramburu, M.J.: Extracting temporal references to assign document event-time periods. In: Proceedings of the 12th International Conference on Database and Expert Systems Applications, Springer Verlag (2001).

[10] Setzer, A.: Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study. PhD thesis, University of Sheffield (2001)

[11] B. Martins, H. Manguinhas, and J. Borbinha. Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. Intl. Conf. on Semantic Computing, 1–9, 2008.

[12] Jannik Strötgen , Extraction and Exploration of Spatio-Temporal Information in Documents.10': Proceedings of the 6th Workshop on Geographic Information Retrieval.

[13] wei wang et al, "Automated spatiotemporal and semantic information extraction for hazards" in journal of Computers, Environment and Urban Systems http://dx.doi.org/10.1016/j.compenvurbsys.2014.11.001 0198-9715/_ 2014 Elsevier Ltd.

[14] David O'Steen et al, 'Named Entity Recognition in Arabic: A Combined Approach' June 4, 2009 Final Project .CS 224N / Ling 237.

[15] KHALED SHAALAN ET AL 'NERA: NAMED ENTITY RECOGNITION FOR ARABIC ' JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY' (IMPACT FACTOR: 2.23). 08/2009; 60(8). DOI: 10.1002/ASI.21090.

[16] Abdulgabbar Mohammad Saif et al,' An Automatic Collocation Extraction from Arabic Corpus' Journal of Computer Science 7 (1): 6-11, 2011 ISSN 1549-3636 © 2011 Science Publications.

[17] Oudah, M., & Shaalan, K. F. (2012). A Pipeline Arabic Named Entity Recognition using a Hybrid Approach. In COLING (pp. 2159-2176).

[18] Zayed, O.H., El-Beltagy, S.R et al .: Person Name Extraction from Modern Standard Arabic or Colloquial Text. In: Proceedings of the eighth (08) International Conference on Informatics and Systems, INFOS 2012, pp. NLP-44–NLP-48.Egypt (2012)

[19] Hala Elsayed et al 'Information Extraction from Arabic News' in IJCSI International Journal of Computer Science Issues, Volume 12, Issue 1, No 2, January 2015 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.

[20] Cohen, S.: Entity extraction enables "discovery". Technical report, Basis Technology (2006).

[21] Technologies, B.: BBN IdentiFinder Text Suite [Online; accessed 22-April- 2010].

[22] COLTEC: Anee: Arabic named entity extraction. Technical report, Computer & Language Technology (2007).

[23] GATE: http://gate.ac.uk/

[24] Fadi zaraket et al 'Arabic Temporal Entity Extraction using Morphological Analysis' in IJCLA vol. 3, no. 1, jan-jun 2012, PP. 121–136 received 29/10/11 Accepted 09/12/11 final15/06/12.

[25] SAMA: http://catalog.ldc.upenn.edu/LDC2010L01.

[27] Roche, M., Heitz, T., Matte-Tailliez, O., Kodratoff, Y.et al , EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés, 2004.

[28] H.Cunningham, D.Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework & Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), 2002.
[29] Plamondon, L., 2004, L'ingénierie de la langue avec GATE, RALI/DIRO, Université de Montréal
[30] http://tv.echoroukonline.com/
[31] http://www.elkhabar.com/

AUTHORS

Abdelkoui feriel (Computer Science), is a Phd student at the University of MENTOURI 2, MISC Laboratory, Her research interests include: GIR, spatial ontologies, machine learning, Data Mining and Arabic language processing.

And M.Kholladi Mohamed khireddine (Computer Science) is a Professor and Rector of HAMMA lakhdar, El-oued University, and the president of MISC Laboratory.