International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244
Impact Factor: 4.054

SETHEMO: THEMATIC SEGMENTATION-BASED ONTOLOGY

Rachid Boudouma*

Abstract: *In this paper, we suggest our automatic system SeThemO (Thematic Segmentation-based Ontology) which allows, through the integration of domain ontology, topic borders detection in a specific textual document. We have implemented it on basis of the Conceptual Energy of Auto Associative Memory [Boudouma 13a]. Built with a nested architecture, our system includes seven main modules that we will fully develop in the rest of this article. We also present the performance achieved by the thematic segmentation process SeThemO. We compare it with the results obtained by some known approaches, according to the assessment protocol that we will explain later in this work.*

Keywords: SeThemO, Thematic Segmentation, Domain Ontology, Conceptual Energy, neural *networks, Hopfield networks, topic segmentation.*

*LASTID, Univ of IBN TOFAIL, Kenitra, Morocco.

1. INTRODUCTION

The topic segmentation called also topic boundaries detection aims to locate the theme changes in textual documents [Bestgen 09]. This task presents no great interest outside the processes where it can be integrated, for example in information retrieval, automatic summary generation, broadcast news...

For ten years, several works of thematic segmentation were proposed. The majority of them use mathematical and statistical heuristics, especially what is called, in literature, the lexical cohesion study which is based on the analysis of the repetition of words in the text, for example in [Choi 00], [Ferret 06], [Hearst 97], [Utiyama 01], [Fernández 07] and [Labadié 09]. It then attaches to find the points where the similarity value presents important variations interpreted as failure of the topic continuity.

On the other hand, there is, at rare trend, a completely different approaches based on linguistic rules. These ones consider the speech not as a simple sequence of sentences, but an elaborate structure. It aims to study the coherence of discourse, using rhetorical and discourse analysis, in order to identify text segments pertinently consistent.

A third type of approaches combines statistical rules and linguistic analysis in order to improve performance.

In context of this latter case, we suggest our automatic system *SeThemO* (Thematic Segmentation-based Ontology) [Boudouma 13b] implemented on basis of the *Conceptual Energy* [Boudouma 13a] which allows, through integration of domain ontology and discursive analysis, to segment thematically specific text.

With a modular architecture, *SeThemO* implements the various components of our approach *Conceptual Energy*, in particular the import of ontology and the corpus to segment, the choice of thematic correlation threshold and the options to use in the thematic segmentation process.

In the remainder of this paper we present firstly an overall modular architecture of *SeThemO* and then we detail the principle operation and heuristic bases of each module in the system.

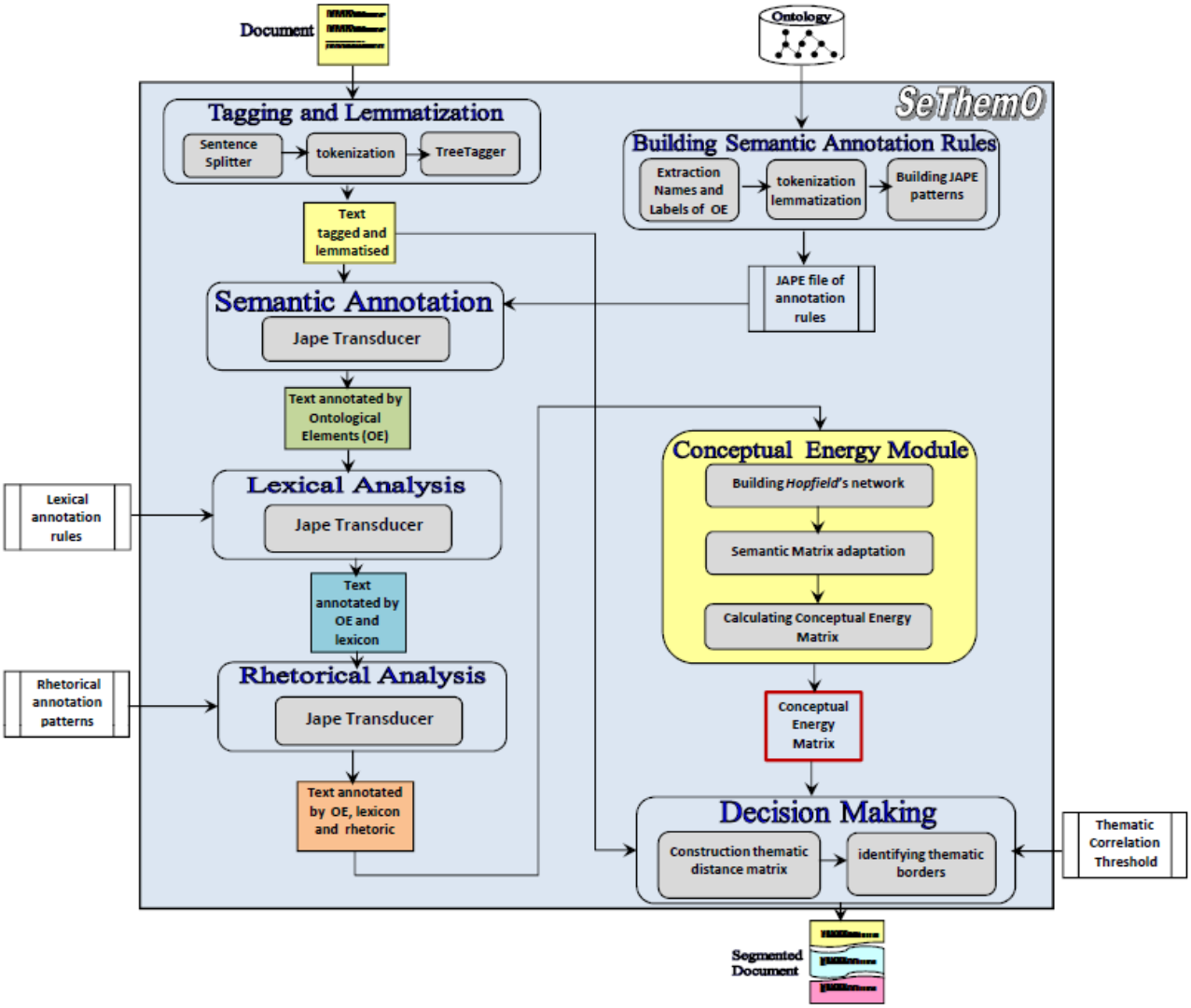
In the last part we illustrate the performance obtained by the thematic detection process *SeThemO* compared with the scores provided by some known algorithms (*TextTiling* [Hearst 97], *C99* [Choi. 00] and *EnerTex* [Fernadez 08]), according to the assessment protocol.

ISSN: 2278-6244
Impact Factor: 4.054

2. FUNCTIONAL ARCHITECTURE OF *SETHEMO*

SeThemO is an automatic system for thematic boundaries detection in the text. It allows entering domain ontology and setting a value of *thematic correlation threshold* to finally have a segmented text.

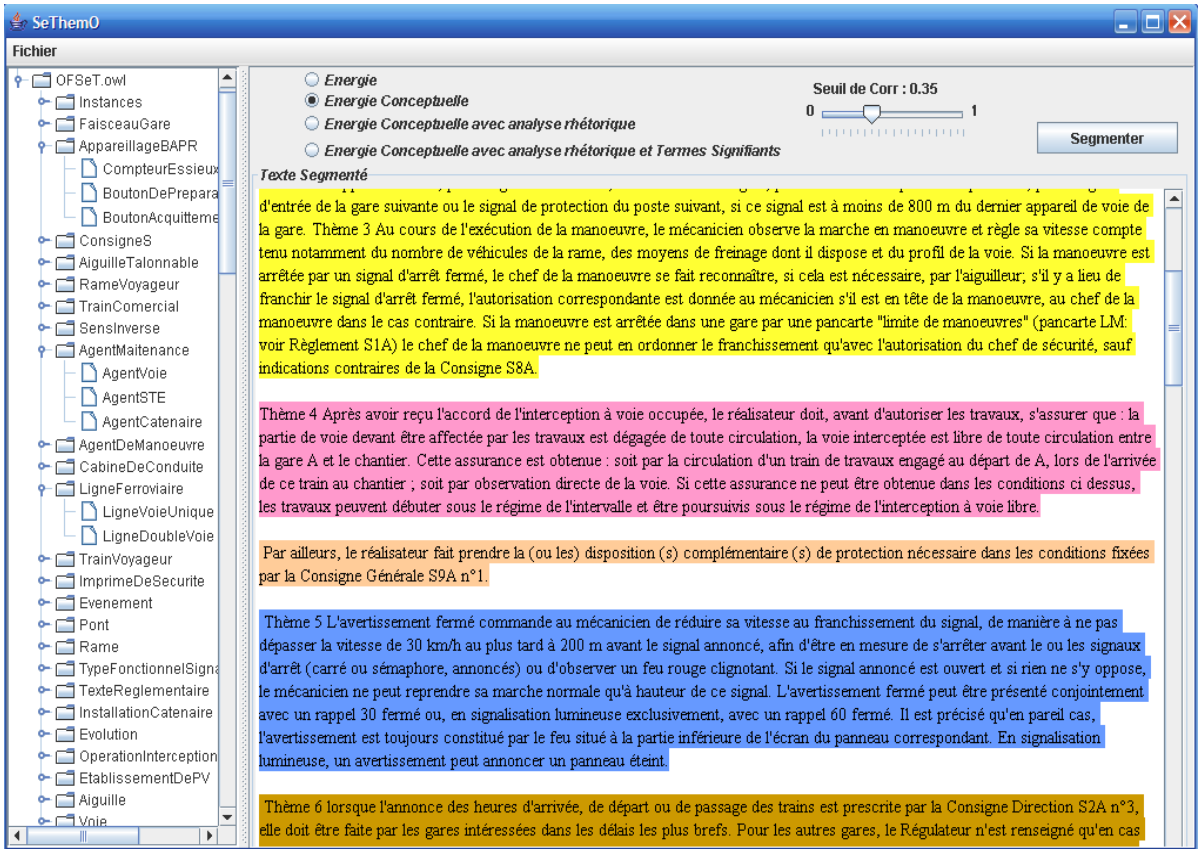
SeThemO has been implemented as a nested architecture as shown in the following figure:



SeThemO architecture

Implemented in Java and using a large number of its technology, *SeThemO* includes seven main functional modules. It allows through its GUI to visualize the segments obtained in different background colors as shown in the below screenshot:

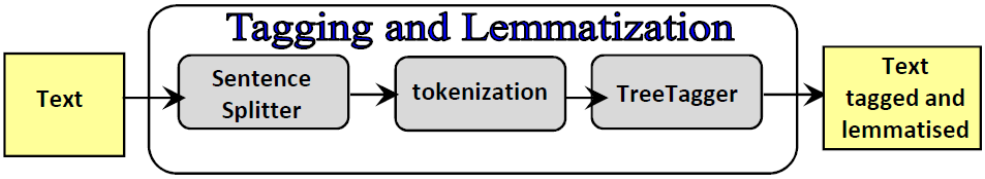
ISSN: 2278-6244
Impact Factor: 4.054



SeThemO GUI

3. BASIC MODULES AND HEURISTICS OF *SETHEMO*

3.1. TEXT TAGGING AND LEMMATIZATION

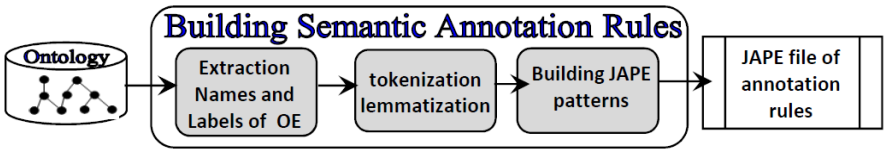


Module of Tagging and Lemmatization

This module takes as input a plain text document and uses *TreeTagger* [Schmid 94] as a grammatical tagger of the text for lemmatizing the terms and determining their grammatical categories in order to reduce the number of morphological variances that can be found in the processed text.

This module includes also the GATE [Cunningham 02] operations "Sentence Splitter" and "tokenization" that consist respectively of dividing the text into sentences and identifying the basic linguistic entities (token, punctuation ...).

3.2. BUILDING SEMANTIC ANNOTATION RULES



Module of building of semantic annotation rules

It is an algorithm that uses JENA technology to extract from the ontology the useful meta-data (class names, labels, attribute names, relations names, and instance names) for building automatically the JAPE (Cunningham et *al.,* 00) rules used for annotating ontological elements co-occurrences in the text.

After questioning the ontology to retrieve the list of the ontological elements (OE), the algorithm through this list and for each OE:

1. Get the name of the OE to use as a rule name. This name is used in the header and in the end of the rule ;
2. Built the header and the end of the rule ;
3. For the same OE, gets each '*label*' and starts the construction of the rule body
4. During the construction of the rule body, the words that constitute the label are lemmatised before being used.

Algorithm 1. Building JAPE rules

```
1: OE : Ontological Element
2: OE : Ontological Elements list
3: rules ← "";
4: For each OE ϵ OEs do {
5:      ruleName ← getNameOf(OE) ;
6:      listLabels ← getLabelsOf(OE)
7:      ruleHeader ← "Rule: "+ ruleName(OE) + "(";
8:      endOfRule ← "):"+ ruleName +"-->:" + ruleName + "." + ruleName + "={kind="" + rule-
                Name +"", rule="+ ruleName + "}";
9:      ruleBody ← "";
10:     For each   label ϵ listLabels   do {
11:          labelBady ← "(";
12:          For each   word ϵ label   do {
13:               wordLem ← lemmeOf(word);
14:               labelBody ← labelBody + "Token.lemma==""+ wordLem +"" ";
15:          }//endFor
16:          if label is not the last {
17:               labelBody ← labelBody +") |"
18:               Else labelBody ← labelBody +")"
19:          }//endif
20:          ruleBody ← ruleBody + labelBody
```

```
21:     } //endFor
22:     rules ← rules + ruleHeader + ruleBody + endOfRule;
23: } //endFor
24: //end
```

The example below gives an extract of *OFSeT* ontology [Boudouma 13a] concerning the concept '*AgentTrain*' expressed in OWL language:

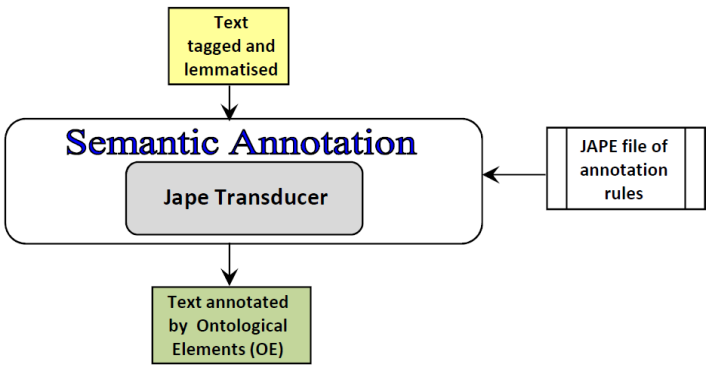
```
1: <!--htt p://www:semanticweb:org=ontologies/2010/0/20/OFSeT:owl#AgentTrain-->
2: < owl : Class rdf : about = "&ontologies;OFSeT:owl#AgentTrain" >
3:                 < rdfs : label xml : lang = "fr" > agents de trains < /rdfs : label >
4:                 < rdfs : label xml : lang = "fr" > agents des trains < /rdfs : label >
5:                 < rdfs : label xml : lang = "fr" > agents du service des trains < /rdfs : label >
6:                 < rdfs : label xml : lang = "fr" > brigade de conduite < /rdfs : label >
7:                 < rdfs : label xml : lang = "fr" > personnel de conduite < /rdfs : label >
8:                 < rdfs : label xml : lang = "fr" > personnel des trains < /rdfs : label >
9:                 < rdfs : label xml : lang = "fr" > personnel train < /rdfs : label >
10:                < rdfs : subClassOf rdf : resource = "&ontologies;OFSeT:owl#Agent"
11: / >
```

The JAPE rule produced for this example by the algorithm 1 will have the following form:

Algorithm 2. JAPE rule 'AgentTrain'

```
1: Rule : AgentTrain
2:   (
3:     (Token:lemma == "agent" Token:lemma == "de" Token:lemma == "train")
4:     | (Token:lemma == "agent" Token:lemma == "du" Token:lemma == "train")
5:     | (Token:lemma == "agent" Token:lemma == "du" Token:lemma == "service"
                    Token:lemma == "du" Token:lemma == "train")
6:     | (Token:lemma == "brigade" Token:lemma == "de" Token:lemma == "conduite")
7:     | (Token:lemma == "personnel" Token:lemma == "de" Token:lemma == "conduite")
8:     | (Token:lemma == "personnel" Token:lemma == "du" Token:lemma == "train")
9:     | (Token:lemma == "personnel" Token:lemma == "train")
10:  ) : AgentTrain-- >: AgentTrain:AgentTrain = kind = "AgentTrain"; rule = AgentTrain
```

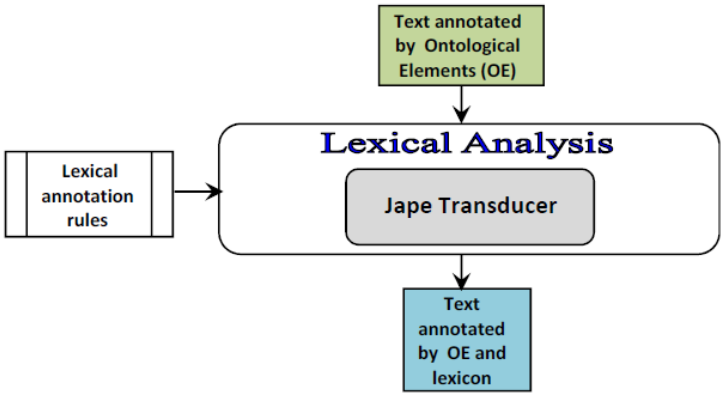
3.3. SEMANTIC ANNOTATION



Module of semantic annotation

This module is responsible of the research and the semantic referencing of the various linguistic forms that are matched by the rules built in the previous phase, using "JAPE transducer" proposed in GATE. We get in the output, a text annotated by various ontological elements.

3.4. LEXICAL ANALYSIS

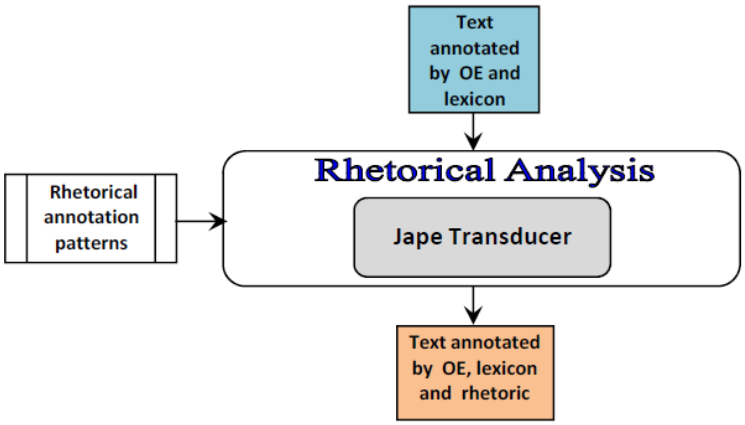


Module of lexical analysis

This step allows annotating, through a JAPE rule, the significant terms which have not been already annotated by the Semantic annotation module.

In the output of this process we have a text containing all the annotations related to OE and meaningful lexicon.

3.5. RHETORICAL ANALYSIS



Module of rhetorical analysis

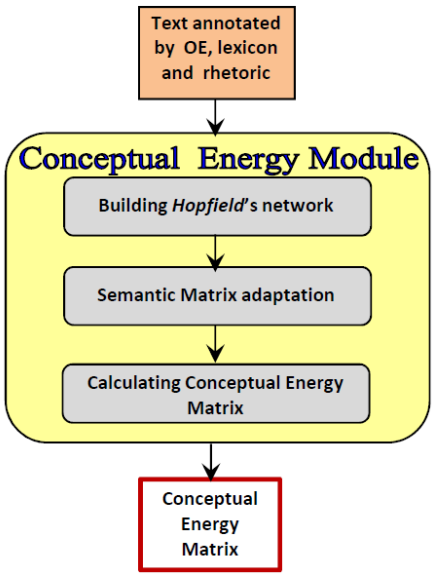
The module consists of locating some connectors and anaphora inspired from *LEXCONN* [Roze 10] by using JAPE Transducer.

ISSN: 2278-6244
Impact Factor: 4.054

Indeed, it takes as input the text resulting from the previous phases and uses a JAPE rule which implements the patterns related to the above connectors and anaphora.

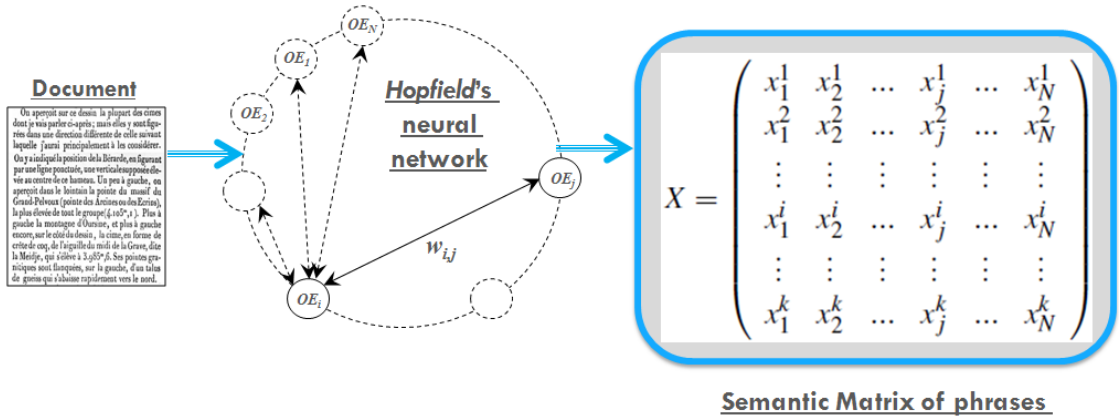
Finally we will have in the output the text annotated by OE, lexicon and rhetoric markers.

3.6. CONCEPTUAL ENERGY MODULE



Conceptual Energy module

This module uses the text with the various annotations resulting from the previous modules in order to build the *Hopfield*'s network [Hopfield 82] and calculate the various matrices, including the Conceptual Energy matrix.



Semantic Matrix of phrases

OE_j Neuron representing the Ontological Element *j*

k : number of sentences the text

N : dimension of the vector space of *OE* contained in text

x_j^i : magnitude value of the neuron representing OE_i relative to the sentence *j*

ISSN: 2278-6244
Impact Factor: 4.054

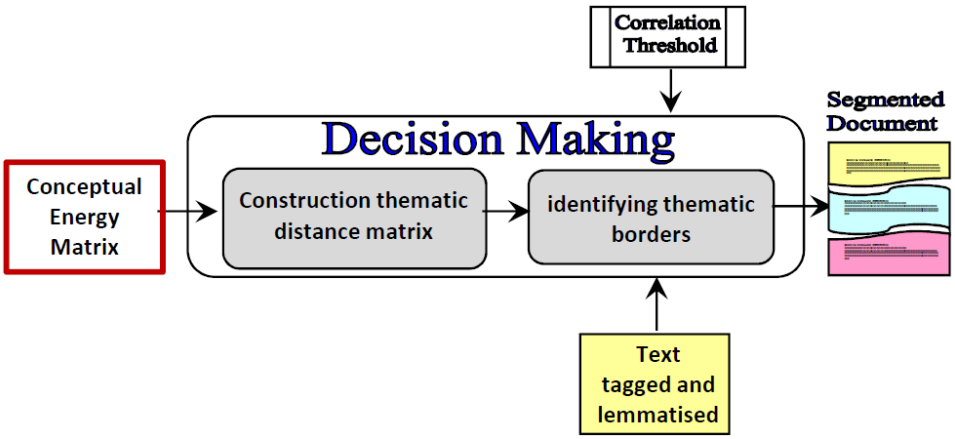
Calculating Conceptual Energy Matrix

$$E = -\frac{1}{2} \times W^2$$

Where *W* is the Semantic correlation matrix calculated under the following formula:

$$W = X \times X^T$$

3.7. DECISION MAKING



Decision making module

This artefact takes as input the *Conceptual Energy* matrix and the value of the *thematic correlation threshold* (*tct)* for calculating the thematic distance matrix. We do this by converting the matrix through the coefficient *KANDELL*.

Coefficient *KENDALL*

$$E = \begin{pmatrix} e_1^1 & e_2^1 & \dots & e_j^1 & \dots & e_k^1 \\ e_1^2 & e_2^2 & \dots & e_j^2 & \dots & e_k^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ e_1^i & e_2^i & \dots & e_j^i & \dots & e_k^i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ e_1^k & e_2^k & \dots & e_j^k & \dots & e_k^k \end{pmatrix} \qquad Dth = \begin{pmatrix} \tau_1^1 & \tau_2^1 & \dots & \tau_j^1 & \dots & \tau_k^1 \\ \tau_1^2 & \tau_2^2 & \dots & \tau_j^2 & \dots & \tau_k^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \tau_1^i & \tau_2^i & \dots & \tau_j^i & \dots & \tau_k^i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \tau_1^k & \tau_2^k & \dots & \tau_j^k & \dots & \tau_k^k \end{pmatrix}$$

τ_j^i coefficient *KENDALL* value between *Ei* and *Ej* that are respectively the Energy Conceptual vector of the i^{th} and the j^{th} sentences

Finally, according to the *tct* chosen, this module specifies the topic boundaries (sentences *i*), using the condition $Dth(i, i+1) < tct$ where $(0 < tct < 1)$. .

EVALUATION OF *SETHEMO*

3.8. EVALUATION PROTOCOL

So that we can evaluate our system, we used the *OFSeT* ontology built in a preceding work [Boudouma 13a] that models the railway safety domain.

We used a test text constituted by concatenated paragraphs dealing with different topics of the domain; this corpus has been prepared by the railway domain experts who have specified manually the thematic borders.

Table 1. ***Text Charactéristics***

Words Number	Sentences Number	Thematics Number	Thematic borders
4068	98	21	3, 6, 9, 13, 18, 26, 31, 37, 42, 45, 49, 52, 56, 58, 61, 65, 67, 72, 79, 84 et 91

The evaluation focuses on the comparison of the results obtained under optimum conditions, by the algorithms *TextTiling*, *C99* et *EnerTex*. These results are expressed as conventional indicators (*recall, precision, F-score* and *WindowDiff*)

3.9. EVALUATION RESULTS

This benchmarking was performed on the same test corpus and with the optimal parameters of each algorithm. The optimal result of *EnerTex* was achieved for 0.70 of *tct* (*thematic correlation threshold).* As for *SeThemO,* it generated its maximum results with a value of 0.43. On *C99* and *TextTiling,* we obtained results by varying their algorithm parameters (*Window size* and *Size of ranking mask*).

The experimental results are summarized in the table below:

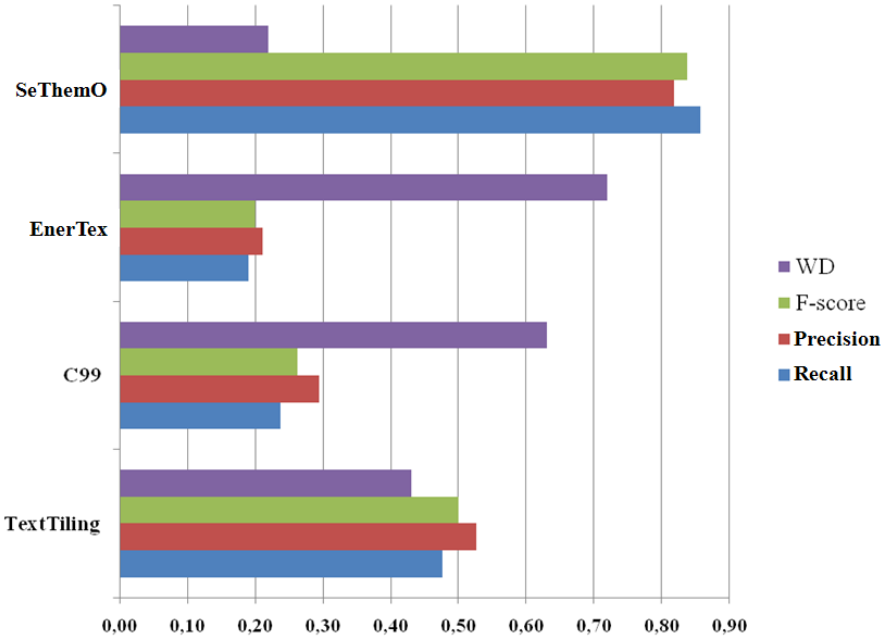
Table 2. ***RESULTS OF EXPERIENCE***

	TextTiling	***C99***	***EnerTex***	***SeThemO***
Recall	0,48	0,24	0,19	**0,86**
Precision	0,53	0,29	0,21	**0,82**
F-score	0,50	0,26	0,20	**0,84**
WD	0,43	0,63	0,72	**0,22**

We observe that *SeThemO* gives the best results for all markers used (86% *recall*, 82% *precision*) which is very distant from *TextTiling* (2nd best performing algorithm 48% *recall*, 53% *precision*).

The same finding was recorded in terms of *windowDiff* with a rate of 0.22 against 0.43 of *TextTiling.*

The comparisons of these results are graphically visualized in the following figure:



4. CONCLUSION

In this work we presented *SeThemO* an automatic system of thematic boundaries detection in a specific text. Thus, we implemented it on the basis of the approach of *Energy Conceptual* [Boudouma 13a]. So by exploiting the domain ontology, *SeThemO* offers an alternative against the algorithms that use the lexicon.

We bring out the system architecture, the operating mechanisms of its various modules as well as its basic technologies and heuristics.

The evaluation of our system, on a test corpus of railway domain, has shown interesting results against some of the most popular algorithms. However, the effectiveness of *SeThemO* in other domain other than the railways must be confirmed.

REFERENCES

1. [Bestgen 09] Bestgen Yves (2009). n Quel indice pour mesurer l'efficacité en segmentation de textes? z. Proceeding of TALN 2009, Senlis, 24-26 juin 2009.
2. [Boudouma 13a] Boudouma Rachid, Raja Touahni et Rochdi Messoussi (2013). New approach for topic segmentation of railway text. In ZENITH: INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH Vol. 3 Issue 2, Feb 2013 issue of ZIJMR HINDIA.

3. [Boudouma 13b]Rachid Boudouma. Raja Touahni et Rochdi Messoussi. *SeThemO* : Système automatique de Segmentation Thématique à base d'Ontologie. In Proceedings of JDTIC 13, Faculté des Sciences de l'Université Ibn Tofail Kénitra, 03- 05 octobre 2013
4. [Cunningham 00] Cunningham. H, D. Maynard and V. Tablan (2000). JAPE: A Java Annotation Patterns Engine, Department of Computer Science, University of Sheffield, 2000.
5. [Cunningham 02] Cunningham H, Maynard D, Bontcheva K, Tablan V (2002). GATE : A framework and graphical development environment for robust NLP tools and applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, US.
6. [Choi 00] Fred Y. Y. Choi. n Advances in domain independent linear text segmentation. z. Proceeding of NAACL-00, pp 26–33, 2000.
7. [Ferret 06] Olivier Ferret. n Approche endogène et exogène pour améliorer la segmentation thématique de documents z. TAL, 2006.
8. [Fernández 07] Énergie textuelle de mémoires associatives. TALN 2007, Toulouse, 5–8 juin 2007
9. [Fernández 08] Fernández S., Sanjuan E. & Torres-Moreno J. M. (2008). Enertex : un système basé sur l'énergie textuelle. In TALN 2008, Avignon, 9-13 juin 2008.
10. [Hearst 97] M. A. Hearst. n TextTiling : Segmenting text into multiparagraph subtopic passages. z. Computational Linguistics, pp 33–64, 1997.
11. [Hopfield 82] Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the USA, 9, 2554–2558.
12. [Labadié 09] Alexandre Labadié Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Français z. Dans thèse pp 5-25, 2009.
13. [Roze 10] Roze C., Danlos L. & Muller P. (2010). LEXCONN : a French Lexicon of Discourse connectives. In Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010), Moissac, France.
14. [Schmid 94] Helmud Schmid. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, p. 44–49.
15. [Utiyama 01] M. Utiyama et H Isahara. n A statistical model for domain independent text segmentation z. ACL, pp 491–498, 2001.