

MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

EXTERLOG : EXtraction de la TERminologie à partir de LOGs

Hassan Saneifar

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM)
Université Montpellier 2 - CNRS UMR 5506

Satin IP Technologies
Cap Omega, RP Benjamin Franklin, 34960 Montpellier Cedex 2, France

Contact : saneifar@lirmm.fr

Résumé

Les fichiers logs issus des systèmes numériques contiennent des informations importantes concernant les conditions et les configurations du système. Dans le domaine de la conception de circuits intégrés, des fichiers logs sont produits par les outils de conception mais ne sont pas systématiquement exploités de façon optimale. Bien que ces logs soient écrits en anglais, ils ne respectent généralement pas la grammaire ou les structures du langage naturel. En outre, ils ont des structures hétérogènes et évolutives. Selon les particularités de telles données textuelles, l'application des méthodes classiques de TALN n'est pas une tâche facile, particulièrement pour extraire la terminologie. Dans cet article, nous présentons notre approche EXTERLOG qui extrait la terminologie à partir des logs. Nous étudions également si l'étiquetage grammatical de fichiers logs est une approche pertinente pour extraire la terminologie.

Abstract

In many domains, the log files generated by digital systems contain important information about the conditions and configurations of systems. In the case of Integrated Circuit designs, log files generated by design tools are not exhaustively exploited. Although these log files are written in English, they usually do not respect the grammar and the structures of natural language. Besides, such logs have a heterogeneous and evolving structure. According to the features of such textual data, applying the classical methods of information extraction is not an easy task, especially for extracting the terminology. In this paper, we thus introduce our approach EXTERLOG to extract the terminology from such log files. We also aim at knowing if POS tagging of such log files is a relevant approach to extract terminology.

Mots-clés : Traitement du langage naturel, Extraction d'information, Extraction de la terminologie, fichiers logs

Keywords: Natural Language Processing, Information Retrieval, Terminology Extraction, Log Files

1. Introduction

Dans de nombreux domaines d'application, les systèmes numériques produisent automatiquement des rapports. Ces rapports produits, appelés logs de système, représentent la source principale d'informations sur la situation des systèmes, des produits ou même des causes des problèmes ayant pu se produire. Dans certains domaines d'application comme les systèmes de conception de Circuits Intégrés, les fichiers logs ne sont pas exploités de façon systématique. Pourtant les fichiers logs générés par les outils de conception contiennent des informations essentielles sur les conditions de production et sur les produits finaux. Pour extraire de l'information à partir des

données textuelles, des méthodes de traitement automatique du langage naturel (TALN) sont classiquement appliquées. Or, ces méthodes ne sont pas adaptées aux caractéristiques spécifiques des données textuelles comme les fichiers logs. Dans ce contexte, le défi principal consiste à fournir des approches qui considèrent des structures et un vocabulaire variables, hétérogènes et évolutifs à partir de telles données textuelles. Bien que le contenu de ces logs ait des similitudes avec les textes écrits en langue naturelle (LN), il ne respecte ni la grammaire du LN ni sa structure. Par conséquent, nous avons besoin d'étudier si les méthodes "classiques" du TALN sont pertinentes dans ce contexte particulier.

Dans cet article, nous avons pour objectif final de créer une ontologie du domaine à partir de ces fichiers logs. Les éléments qui constituent cette ontologie sont les termes issus des logs. Nous étudions ici la pertinence de deux approches principales d'extraction de terminologie à partir de ces fichiers logs. Ces deux approches consistent à extraire des co-occurrences avec et sans utilisation de patrons syntaxiques.

Dans la section 2, nous développons l'utilité des ontologies du domaine et l'extraction de terminologie dans notre contexte ainsi que les particularités et les difficultés de ce contexte. L'approche EXTERLOG : EXtraction de la TERminologie à partir de LOGs est développée dans la section 3. La section 4 explique et compare les différentes expérimentations que nous avons effectuées pour extraire la terminologie à partir des logs.

2. Problématique

Aujourd'hui, les systèmes numériques génèrent de nombreux types de fichiers logs, ce qui donne les informations essentielles sur les systèmes. Certains types de fichiers logs, comme les logs issus des systèmes de surveillance de réseau, des interactions des services web ou des logs d'utilisation des sites web (*web usage*) sont largement exploités [7,9,16]. Ces types de fichiers s'appuient sur la gestion des événements. Cette situation signifie que les informations se trouvant dans ce type de logs sont des événements survenus dans le système qui sont enregistrés chronologiquement. Le contenu de ces logs se conforme à certaines normes grâce à la nature des événements et de leurs utilisations universelles (e.g. les services du web).

Au contraire, dans certains domaines comme les systèmes numériques de conception des circuits intégrés, les logs générés sont plutôt des rapports numériques que l'enregistrement de certains événements. L'objectif de l'exploitation de ces fichiers logs n'est pas d'analyser les événements mais d'extraire de l'information. Par exemple, des configurations de système dans une situation donnée. Ces fichiers logs sont une source importante d'information pour les systèmes d'information conçus pour interroger et gérer la ligne de production.

2.1. Extraction d'information dans les logs

Afin d'utiliser ces logs dans un système d'information, nous devons implémenter des méthodes d'extraction d'information adaptées aux caractéristiques de ces logs. Les caractéristiques particulières de ces données textuelles rendent peu utilisables les techniques statiques fondées sur l'utilisation simple des schémas d'extraction (e.g. expressions régulières) pour extraire l'information pertinente.

Dans la conception des circuits intégrés, il existe plusieurs niveaux. Chaque niveau correspond à une étape de conception. Dans chaque étape, on considère et vérifie certains caractéristiques des CI. Les logs de chaque niveau de conception contiennent des informations sur des caractéristiques propres du niveau. À chaque niveau, plusieurs outils de conception peuvent être utilisés. Malgré le fait que les logs issus du même niveau de conception contiennent les mêmes informations, les structures peuvent significativement différer en fonction de l'outil de conception utilisé. Plus précisément, pour la même information, chaque outil de conception possède souvent son propre vocabulaire. Par exemple, nous faisons produire deux fichiers logs (e.g. log "A" et log "B") par deux outils différents. Une information du type "Statement coverage" sera exprimée sous la forme suivante dans le log "A" :

TOTAL	COVERED	PERCENT	
Lines	10	11	12
statements	20	21	22

Mais cette même information dans le log "B", sera exprimée par cette simple ligne :

```
"EC : 2.1%"
```

Tel que montré ci-dessus, la même information, dans deux fichiers logs produits par deux outils différents, est représentée par des structures et un vocabulaire totalement différents. En outre, l'évolution des outils de conception fait changer le format des données, ce qui rend inefficace l'utilisation des méthodes statiques. Nous avons également observé que plusieurs mots sont utilisés pour le même concept. Pour le concept du temps par exemple, les mots suivants peuvent être trouvés : `Clk`, `CLK`, `Clock`.

Par conséquent, nous avons besoin de méthodes généralisées qui prennent en compte l'hétérogénéité de la structure et du vocabulaire des logs. Comme dans de nombreux travaux [6], pour améliorer et généraliser au mieux les schémas d'extraction, nous avons besoin d'ontologies du domaine qui font la correspondance entre des termes utilisés dans les logs issus des différents outils. Nous utiliserons cette ontologie pour diminuer l'hétérogénéité des termes issus des outils de conception différents. Par exemple, pour vérifier "*l'absence des attributs*" sur les logs, nous devons chercher des phrases différentes dans les logs en fonction de la version et du type d'outil de conception utilisé :

- "**Attribute specifications** are illegal in entities"
- "Do not use **map_to_module attribute**"
- "Do not use **one_cold** or **one_hot attributes**"
- "Do not use **enum_encoding attribute**"
- "The **EVENT attribute** is not supported in subprograms"

Au lieu d'utiliser plusieurs patrons chacun adapté à une phrase, en associant les termes "Attribute specifications", "map_to_module attribute", "one_hot attributes", "enum_encoding attribute" et "EVENT attribute" au concept "*Absence d'attributs*", nous utiliserons un patron général qui s'adapte (*se développe*) automatiquement aux différents logs en utilisant l'ontologie du domaine.

Les méthodes de TALN, notamment celles de l'extraction de la terminologie, développées pour les textes écrits en langue naturelle, ne sont pas forcément bien adaptées aux logs. L'hétérogénéité des données traitées existe non seulement entre les fichiers logs produits par différents outils de conception, mais également au sein d'un fichier logs donné. Par exemple, les symboles utilisés pour présenter la même notion comme l'entête des tableaux changent au sein d'un log. De même, il existe plusieurs formats pour les ponctuations, les lignes de séparation et la représentation de données manquantes. De plus, certains caractères communs sont utilisés pour présenter différentes notions ou structures. En outre, beaucoup de notions dans ces fichiers sont des symboles, des abréviations ou des mots techniques, qui sont seulement compréhensibles en considérant la documentation du domaine. Ainsi, plusieurs lexiques de ce domaine sont constitués de caractères alphanumériques et spéciaux. Par ailleurs, le langage utilisé dans ces logs est une difficulté qui influence les méthodes d'extraction d'information. Bien que la langue utilisée dans ces logs soit l'anglais, les contenus de ces logs n'en respectent pas la grammaire "classique". Dans cet article, nous proposons donc une approche d'extraction de terminologie EXTERLOG adaptée à ces particularités.

2.2. Méthodes d'extraction de la terminologie

L'extraction de la terminologie de domaine dans les données textuelles est une tâche essentielle afin d'établir des dictionnaires spécialisés du domaine [12]. Les bigrammes¹ sont utilisés dans [11] comme des attributs (*features*) pour améliorer la performance de la classification de textes. Les séries de trois mots (*i.e.* trigrammes) ou plus ne sont pas toujours essentiels [8].

Les règles et les grammaires définies sont utilisées dans [4] afin d'extraire les termes nominaux ainsi que pour les évaluer. EXIT, présenté par [12], est une approche itérative qui extrait les termes de façon incrémentale. Un terme découvert à une itération est utilisé dans la prochaine itération afin de trouver des termes plus complexes. Certains travaux tentent d'extraire les co-occurrences dans une fenêtre de taille fixe². Dans ce cas, les mots extraits pour former un terme peuvent ne

1. un n-gramme de mots est défini comme une série de "n" mots.

2. une fenêtre de mots

pas être directement liés [10]. XTRACT évite ce problème en considérant les positions relatives des co-occurrences. XTRACT est un système d'extraction de terminologie, qui identifie des relations lexicales dans les corpus volumineux de textes anglais [15]. SYNTAX, présenté par [1], effectue l'analyse syntaxique des textes pour identifier les noms, les verbes, les adjectifs, les adverbes, les syntagmes nominaux et les syntagmes verbaux. Il analyse les textes en appliquant des règles syntaxiques pour en extraire les termes. TERMEXTRACTOR, présenté par [14], est un logiciel pour l'extraction des termes pertinents dans un domaine spécifique. L'application prend en entrée un corpus de documents de domaine, analyse les documents, et extrait des termes syntaxiquement plausibles. Afin de sélectionner uniquement les termes qui sont pertinents pour le domaine, certaines mesures fondées sur l'entropie sont utilisées. Notons que dans [13] une comparaison des termes extraits avec notre approche et TERMEXTRACTOR est présentée.

Les méthodes statistiques sont généralement utilisées associées à des méthodes syntaxiques pour évaluer la pertinence des candidats terminologiques [3]. Ces méthodes sont fondées sur des mesures statistiques pour valider comme terme pertinent un candidat extrait. Parmi ces mesures, la fréquence d'occurrences des candidats est une notion de base. Or, ces méthodes statistiques ne sont pas pertinentes dans notre contexte. En effet, les approches statistiques peuvent identifier les termes ayant une fréquence élevée, mais ont tendance à omettre des termes peu fréquents [5]. Selon les fichiers logs décrits ci-dessus, la répétition des mots est plus rare. Chaque partie d'un log contient certaines informations indépendantes d'autres parties. En outre, il n'est pas raisonnable de constituer un corpus volumineux de certains logs générés par le même outil au même niveau de conception.

Beaucoup de travaux comparent les différentes techniques d'extraction de terminologie et leur performance. Mais la plupart de ces travaux sont expérimentés sur les données textuelles, qui sont les textes classiques écrits en langage naturel. La plupart des corpus utilisés sont structurés de manière cohérente. En particulier, ces données textuelles suivent la grammaire du LN. Or, dans notre contexte, les caractéristiques particulières des logs imposent une adaptation particulière pour que ces méthodes soient pertinentes.

3. EXTERLOG : EXtraction de la TERminologie à partir de LOGs

Dans cette section, nous allons détailler notre approche EXTERLOG afin d'extraire la terminologie dans les fichiers logs. Notre processus consiste dans un premier temps à la normalisation des fichiers logs et à leur étiquetage grammatical. Cette préparation des données permet alors d'extraire des bigrammes selon des patrons syntaxiques.

3.1. Normalisation

Compte tenu des spécificités de nos données textuelles, nous appliquons une méthode de normalisation adaptée aux logs pour rendre le format et la structure des logs plus cohérents. Nous remplaçons les ponctuations, les lignes de séparation et les en-têtes des tableaux par des caractères spéciaux pour réduire l'ambiguïté. Puis, nous segmentons les mots, considérant que certains mots ou structures ne doivent pas être segmentés. Par exemple, le mot technique "Circuit4-LED3" est un mot unique qui ne doit pas être segmenté en deux mots "Circuit4" et "LED3". De plus, nous identifions les lignes représentant l'en-tête des tableaux pour distinguer les lignes de séparation. Cette normalisation rend la structure des logs issus des différents outils plus homogène.

3.2. Étiquetage grammatical

L'étiquetage grammatical est une méthode classique du TALN pour réaliser l'annotation grammaticale des mots. Dans notre contexte, selon la nature des fichiers logs, il existe des difficultés et des limites pour appliquer un étiqueteur grammatical sur de telles données textuelles. Pour identifier le rôle grammatical des mots dans les logs, nous employons l'étiqueteur grammatical BRILL [2]. Nous avons adapté BRILL au contexte des logs en utilisant de nouvelles règles *contextuelles* et *lexicales*. Par exemple, un mot commençant par un nombre est considéré comme un "cardinal" par l'étiqueteur BRILL. Or, dans les fichiers logs, il existe de nombreux mots comme 12.1vS010; qui ne devraient pas être étiquetés comme "cardinal". C'est pourquoi nous avons défini des règles lexicales et contextuelles spécifiques à nos données.

Nous maintenons la structure des textes inchangée lors de l'étiquetage des logs car elle peut être

une information importante pour extraire l'information dans nos travaux futurs. Pour cela, nous présentons de nouvelles étiquettes comme "\TH", qui représente la ligne utilisée dans l'en-tête des tableaux ou "\SPL" pour les lignes de séparation. Ces étiquettes, que nous appelons "*les étiquettes de structure de document*", peuvent être identifiées lors de la normalisation et par des règles contextuelles/lexicales spécifiques qui ont été introduites dans BRILL. Le corpus étiqueté peut alors être utilisé dans la phase suivante du processus qui consiste à extraire la terminologie.

3.3. Extraction des bigrammes

Afin d'identifier les co-occurrences dans les logs, nous considérons deux solutions :

1. extraction des bigrammes en utilisant des patrons syntaxiques (ci-après "*Bigrammes-AP*"³),
2. extraction des bigrammes sans utilisation de patrons syntaxiques (ci-après "*Bigrammes-SP*"⁴).

Dans la première, nous utilisons le filtrage de mots par des patrons syntaxiques. Les patrons syntaxiques déterminent les mots adjacents ayant les rôles grammaticaux définis. Les patrons syntaxiques que nous utilisons pour extraire les bigrammes-AP sont :

- "\AJ - \NN" (Adjectif-Nom)
- "\NN - \NN" (Nom-Nom)

Au contraire, dans la deuxième solution, l'extraction des bigrammes-SP (sans utilisation des patrons syntaxiques) ne dépend pas du rôle grammatical des mots. Afin d'extraire les bigrammes-SP significatifs, nous considérons le bruit existant dans les logs. Par conséquent, nous normalisons et segmentons les logs pour diminuer le taux de bruit. Les bigrammes extraits représentent deux mots adjacents "ordinaires". Dans la section 4, nous analysons les candidats terminologiques obtenus par chacune des méthodes.

4. Expérimentations

Nous avons expérimenté deux approches différentes de l'extraction de terminologie à partir des logs : (1) en utilisant des patrons syntaxiques (*bigrammes-AP*) et (2) sans utilisation des patrons syntaxiques (*bigrammes-SP*). Ici, nous analysons les candidats terminologiques obtenus. Le corpus de logs d'une taille de 950 Ko est constitué des logs de tous les niveaux de conception.

4.1. Termes communs

Nous avons créé la liste globale des bigrammes-AP ainsi que des bigrammes-SP pour chaque outil de conception (dans notre cas les outils "A" et "B") à chaque niveau de conception. Nous comparons ces listes pour déterminer si les logs issus de différents outils partagent les mêmes termes. Malgré la structure totalement différente des logs, nous constatons qu'ils utilisent parfois une terminologie commune. Ce vocabulaire commun peut être employé pour généraliser les pa-

Outils	Niveau 1		Niveau 2		Niveau 3		Niveau 4		Niveau 5	
	A	B	A	B	A	B	A	B	A	B
Bigrammes-AP	3,7	16,3	1,7	25	0,5	6,6	0,7	11,1	5,2	7,6
Bigrammes-SP	8,3	8,9	0,3	16,2	0,3	7,5	2	8,3	1,2	4,6

TABLE 1 – Bigrammes-AP et les bigrammes-SP communs entre deux fichiers logs générés par deux outils de conception A et B du même niveau de conception.

trons d'extraction d'information à partir des logs dans nos travaux futurs. Le Tableau 1 montre le pourcentage de bigrammes-AP et de bigrammes-SP que chaque outil de conception partage avec l'autre outil pour chaque niveau de conception. Par exemple, au premier niveau de conception, le log issu de l'outil "A" partage seulement 3.7% de ses bigrammes-AP avec le log généré, par l'outil "B". Nous présentons ci-dessous quelques exemples de bigrammes-AP communs trouvés

3. AP : Avec Patron

4. SP : Sans Patron

par notre méthode :

sensitivity list, clock signal, enumeration type, sensitivity list, case expression, clock pin.

Au contraire, il existe aussi des termes non communs comme :

"ruleset hdl_naming", "policy designware", "level shifter", "ruleset rtl", "vhdl port"

Les résultats montrent que les différents outils de conception ne partagent pas les termes techniques, qui forment la majeure partie de ces logs. Selon le tableau 1, l'outil "A" partage généralement moins de termes que l'autre outil. Ceci est dû au fait que les logs issus de l'outil "A" sont généralement plus complets et plus volumineux que les logs de l'outil "B". Par conséquent, nous avons davantage de candidats terminologiques dans les logs de l'outil "A". Le taux faible des termes communs montre l'hétérogénéité des logs et de leur vocabulaire malgré le fait qu'ils contiennent le même type d'information.

4.2. Bigrammes-AP vs Bigrammes-SP

Pour analyser la performance des deux approches choisies pour l'extraction des bigrammes, nous devons évaluer les termes extraits. Pour évaluer de manière automatique leur pertinence, nous comparons les bigrammes-AP et bigrammes-SP aux termes extraits à partir des documents de référence du domaine. Pour chaque niveau de conception des circuits intégrés, nous utilisons certains documents, qui expliquent les principes de la conception et particulièrement les détails des outils de conception. Nous employons ces documents comme "références expertes" dans le cadre d'une validation automatique. En effet, si un terme extrait des logs est utilisé dans les références du domaine, nous pouvons le considérer comme un terme valide du domaine. Pourtant, il existe plusieurs termes propres aux logs surtout les termes techniques qui ne sont pas utilisés dans le corpus de référence. C'est pourquoi une validation par un expert, effectuée dans nos futures travaux, est indispensable pour compléter la validation automatique.

Le corpus de documents de référence est composé d'environ trois documents par niveau de conception. Ces documents sont de taille considérable. Chaque document est constitué d'environ 600 pages. Étant donné que le corpus de référence est constitué des textes écrits en langue standard contrairement aux logs, nous appliquons la méthode classique d'extraction de terminologie pour extraire les termes à partir du corpus de référence. Nous calculons la précision et le rappel

Bigrammes	Niveau 1		Niveau 2		Niveau 3		Niveau 4		Niveau 5	
	AP	SP	AP	SP	AP	SP	AP	SP	AP	SP
Précision	67,7	11,3	20,7	6,5	37,8	9,9	40,1	6,5	19,6	5,1
Rappel	0,7	0,4	7,6	7,5	1,3	1,0	9,5	8,8	0,3	0,5

TABLE 2 – Précision et le rappel des bigrammes-AP et des bigrammes-SP selon les termes de référence.

des bigrammes-AP et SP. Dans notre contexte, la précision est calculée comme le pourcentage des bigrammes extraits qui existent dans les termes de référence. Nous calculons le rappel comme le pourcentage des termes de référence qui sont couverts par les bigrammes extraits à partir des logs.

$$\text{Précision} = \frac{|\text{Bigrammes} \cap \text{Termes de référence}|}{|\text{Bigrammes}|}$$

$$\text{Rappel} = \frac{|\text{Bigrammes} \cap \text{Termes de référence}|}{|\text{Termes de références}|}$$

Le tableau 2 montre la précision et le rappel des bigrammes-AP et bigrammes-SP. Pour évaluer les termes extraits des logs, la précision est la mesure la plus adaptée à notre contexte. En effet, cette mesure donne une tendance générale quant à la qualité des termes extraits par notre système. Notons que pour calculer une mesure de précision parfaitement adaptée, il faudrait demander à un expert d'évaluer manuellement tous les termes proposés par EXTERLOG. Or, une telle tâche est

difficile et coûteuse à mettre en œuvre. Par conséquent nous calculons la précision en fonction des termes de référence.

La comparaison des bigrammes-AP et bigrammes-SP relativement aux termes de référence montre que l'extraction de la terminologie fondée sur les patrons syntaxiques est tout à fait pertinente sur les données logs. La précision des bigrammes-AP se révèle en effet plus élevée par rapport aux bigrammes-SP. Malgré le fait que la normalisation et l'étiquetage des données logs ne soient pas une tâche facile, nos expérimentations montrent qu'un effort dans ce sens est tout à fait utile dans le but d'extraire une terminologie de qualité.

Le rappel faible des candidats terminologiques est dû au grand nombre de termes de référence. Le corpus de référence est environ 5 fois plus volumineux que le corpus de logs. En outre, nous avons constaté que beaucoup de candidats terminologiques extraits qui n'ont pas été validés sont des mots ou des abréviations techniques, qui se trouvent seulement dans les logs et non dans les documents de références du domaine. C'est pourquoi les résultats de rappel ne sont pas réellement représentatifs pour évaluer la qualité d'EXTERLOG.

4.3. Élagage des candidats terminologiques

Dans le but de favoriser les candidats terminologiques les plus pertinents et significatifs, nous filtrons les bigrammes extraits en fonction de leurs fréquences d'occurrences dans les logs.

Dans nos données, la fréquence de 66% des termes extraits est égale à 1. Par ailleurs, 21% des termes sont répétés deux fois. Ainsi, le nombre des termes fréquents est très faible. Nous avons alors choisi d'extraire des bigrammes ayant une fréquence au moins égale à "2". Le nombre des candidats terminologiques extraits avant et après élagage est présenté dans le tableau 3.

Bigrammes	Niveau 1		Niveau 2		Niveau 3		Niveau 4		Niveau 5	
	AP	SP	AP	SP	AP	SP	AP	SP	AP	SP
Avant	58	235	2514	8047	264	1040	71	326	205	878
Après	11	55	514	2852	142	531	24	108	78	351

TABLE 3 – Le nombre des candidats terminologiques extraits à chaque niveau avant et après élagage à 2 (suppression des candidats présents une seule fois).

Le tableau 4 montre la précision et le rappel des bigrammes extraits après l'élagage des bigrammes. Les résultats montrent que la précision augmente globalement même si cette dernière diminue

Bigrammes	Niveau 1		Niveau 2		Niveau 3		Niveau 4		Niveau 5	
	AP	SP	AP	SP	AP	SP	AP	SP	AP	SP
Précision	81,1	10,1	18	5	37,2	5,9	27,3	7,1	37,1	5,5
Rappel	0,1	0,1	3	2	0,1	0,4	1,6	2,2	0,2	0,1

TABLE 4 – Précision et rappel des bigrammes-AP et des bigrammes-SP après élagage des bigrammes ayant la fréquence très basse.

dans certains cas. Nous constatons toutefois que l'élagage des termes ayant une fréquence faible n'améliore pas les résultats de manière significative. Comme nous l'avons relevé en section 2.2, les techniques fondées sur la fréquence de termes (élagage selon la fréquence) ne sont pas pertinentes dans notre contexte. En effet, l'élagage risque de supprimer certains termes intéressants qui ont une fréquence faible. Ceci est le cas des termes extraits des logs issus du 2^{ème} niveau de conception.

5. Conclusion & Perspectives

Dans cet article, nous avons décrit un type particulier de données textuelles : fichiers logs générés par des outils de conception de circuits intégrés. Ces fichiers logs ne suivent pas la grammaire du langage naturel. En outre, ils ont des structures très hétérogènes. Pour extraire la terminologie du domaine, nous extrayons des bigrammes avec deux approches différentes : (1) extraction des co-occurrences selon les patrons syntaxiques et (2) extraction sans patrons syntaxiques. Bien que ces textes (*logs*) ne respectent généralement pas la grammaire de la langue naturelle et malgré leurs structures spécifiques, les résultats d'expérimentations montrent que les termes obtenus en utilisant les patrons syntaxiques sont plus pertinents que ceux obtenus sans utilisation de patrons syntaxiques.

Afin d'améliorer la performance de l'extraction de la terminologie, nous allons appliquer d'autres méthodes de normalisation. Étant donnée l'importance de la précision de l'étiquetage, une amélioration de l'étiqueteur grammatical sera effectuée. L'expertise manuelle des termes extraits avec notre système devra être menée afin de confirmer les résultats présentés dans cet article. Enfin, nous envisageons de prendre en compte la terminologie extraite avec notre système afin d'enrichir les patrons d'extraction d'information actuellement utilisés.

Bibliographie

1. Didier Bourigault et Cécile Fabre. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire - Université Toulouse le Mirail*, (25) :131–151, 2000.
2. Eric Brill. A simple rule-based part of speech tagger. In *In Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155, 1992.
3. Béatrice Daille. Conceptual structuring through term variations. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 9–16, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
4. Sophie David et Pierre Plante. De la nécessité d'une approche morpho-syntaxique en analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec*, 2(3) :140–155, September 1990.
5. David A. Evans et Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 1996. Association for Computational Linguistics.
6. Fabrice Even et Chantal Enguehard. Extraction d'informations à partir de corpus dégradés. In *Proceedings of 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'02)*, pages 105–115, 2002.
7. Federico Michele Facca et Pier Luca Lanzi. Mining interesting knowledge from weblogs : a survey. *Data Knowl. Eng.*, 53(3) :225–241, 2005.
8. Marko Grobelnik. Word sequences as features in text-learning. In *In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98)*, pages 145–148, 1998.
9. Dong (Haoyuan) Li, Anne Laurent, et Pascal Poncelet. Mining unexpected web usage behaviors. In *ICDM*, pages 283–297, 2008.
10. Dekang Lin. Extracting collocations from text corpora. In *In First Workshop on Computational Terminology*, pages 57–63, 1998.
11. Chade meng Tan, Yuan fang Wang, et Chan do Lee. The use of bigrams to enhance text categorization. In *Inf. Process. Manage*, pages 529–546, 2002.
12. Mathieu Roche, Thomas Heitz, Oriane Matte-Tailliez, et Yves Kodratoff. EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT'04 (International Conference on Statistical Analysis of Textual Data)*, volume 2, pages 946–956, 2004.
13. Hassan Saneifar, Stéphane Bonniol, Anne Laurent, Pascal Poncelet, et Mathieu Roche. Terminology extraction from log files. In *DEXA'09 : Proceedings of 20th International Conference on Database and Expert Systems Applications*, Lecture Notes in Computer Science, pages 769–776. Springer, 2009.
14. Francesco Sclano et Paola Velardi. Termextractor : a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*, Funchal, Portugal, 2007.
15. Frank Smadja. Retrieving collocations from text : Xtract. *Comput. Linguist.*, 19(1) :143–177, 1993.
16. Kenji Yamanishi et Yuko Maruyama. Dynamic syslog mining for network failure monitoring. In *KDD '05 : Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 499–508, New York, NY, USA, 2005. ACM.