Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 10/2 (2017), 59-66
DOI: http://dx.doi.org/10.21609/jiki.v10i2.374

AUTOMATIC ONTOLOGY CONSTRUCTION USING TEXT CORPORA AND ONTOLOGY DESIGN PATTERNS (ODPS) IN ALZHEIMER'S DISEASE

Denis E. Cahyani[1] and Ito Wasito[2]

[1]Departement of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Jl. Ir. Sutami No.36A, Jebres, Kota Surakarta, Jawa Tengah 57126, Indonesia.
[2]Faculty of Computer Science, Universitas Indonesia, Kampus UI, Depok, 16424, Indonesia

E-mail: denis.eka@staff.uns.ac.id, ito.wasito@cs.ui.ac.id

Abstract

An ontology is defined as an explicit specification of a conceptualization, which is an important tool for modeling, sharing and reuse of domain knowledge. However, ontology construction by hand is a complex and a time consuming task. This research presents a fully automatic method to build bilingual domain ontology from text corpora and ontology design patterns (ODPs) in Alzheimer's disease. This method combines two approaches: ontology learning from texts and matching with ODPs. It consists of six steps: (i) Term & relation extraction (ii) Matching with Alzheimer glossary (iii) Matching with ontology design patterns (iv) Score computation similarity term & relation with ODPs (v) Ontology building (vi) Ontology evaluation. The result of ontology composed of 381 terms and 184 relations with 200 new terms and 42 new relations were added. Fully automatic ontology construction has higher complexity, shorter time and reduces role of the expert knowledge to evaluate ontology than manual ontology construction. This proposed method is sufficiently flexible to be applied to other domains.

Keywords: *fully automatic, ontology building, ontology design patterns, Alzheimer disease*

Abstrak

Ontologi didefinisikan sebagai spesifikasi eksplisit dari sebuah konseptualisasi, yang merupakan alat penting untuk pemodelan, pembagian, dan penggunaan kembali pengetahuan domain. Namun, konstruksi ontologi dengan tangan merupakan tugas yang rumit dan memakan waktu. Penelitian ini menyajikan metode otomatis untuk membangun ontologi domain bilingual dari pola desain korporat teks dan ontologi (ODPs) pada penyakit Alzheimer. Metode ini menggabungkan dua pendekatan: pembelajaran ontologi dari teks dan sesuai dengan ODP. Ini terdiri dari enam langkah: (i) ekstraksi istilah & hubungan (ii) Pencocokan dengan glosarium alzheimer (iii) Pencocokan dengan pola desain ontologi (iv) Perhitungan skor kesamaan istilah & hubungan dengan ODPs (v) Ontologi bangunan (vi) Evaluasi Ontologi. Hasil ontologi yang terdiri dari 381 istilah dan 184 hubungan dengan 200 istilah baru dan 42 hubungan baru ditambahkan. Konstruksi ontologi otomatis lengkap memiliki kompleksitas yang lebih tinggi, waktu yang lebih singkat dan mengurangi peran pengetahuan ahli untuk mengevaluasi ontologi daripada konstruksi ontologi manual. Metode yang diusulkan ini cukup fleksibel untuk diterapkan pada domain lain.

Kata Kunci: *fully automatic, ontology building, pola desain ontologi, penyakit alzheimer*

1. Introduction

Alzheimer's disease is one of the important issues in the field of public health. In France, there are about 860,000 people affected by Alzheimer's disease. Each year there are 220,000 new cases identified in the country. To prevent the increase in patients suffering from Alzheimer's disease and addressing existing cases, clinical epidemiology training required for health workers. The purpose of the training is to be able to improve the quality of health care workers and to increase the number of health workers who are able to contribute to handle cases that occur in Alzheimer's disease. The improvement of quality medical practice needed to prevent the increase of Alzheimer's disease's patient. One of way to prevent this is educating practitioners in clinical epidemiology. So the number of medicinal practice that can be handle cases about Alzheimer's disease increasing.

Conversely, rapid and efficient decision making is a crucial issue in the public health domain and especially in the Alzheimer's disease domain.

Decision-makers should refer to various experts opinions as they cannot screen themselves all the scientific facts reported in different sources including online scientific literatures and results of clinical trials.

Within this framework, there is a corpora named BiblioDem Digital Library. BiblioDem Digital Library is a critical review of scientific papers related to Alzheimer's disease from a variety of reference journals. Critical analysis of the various articles are reviewed by a domain expert research and integrated into an online bibliographic database called BiblioDem. There are 6-11 relevant papers selected for publication in the monthly magazine named BiblioDemences. BiblioDem contains over 1500 documents which contain title, abstract, critical analysis and the name of experts who carried out the analysis. The paper will be published in BiblioDem if the query contains a term which is shown as the title or abstract in the paper.

Beside using a BiblioDem text corpora, this research also uses ontology design patterns. Ontology design patterns (ODPs) is derivative of the design patterns which used in software engineering. Ontology design patterns is a pattern that makes it possible to identify the design of the ontology structure. Design patterns allow for the regulation of inter-term dependencies so if there is a change in the term it will not affect the other terms. Examples of types of ODPs is extensional patterns, good practice patterns, modeling patterns that can be implemented using the OWL format [1].

This research also uses the Alzheimer glossary to filter word extracted from text corpora BiblioDem. Glossary Alzheimer's is a list of vocabulary and their definitions related to Alzheimer's disease. A glossary contains explanations of concepts relevant to a particular topic and related to the ontology.

This research will be conducted in bilingual domain ontology construction using a text corpus and matching with ontology design patterns for representing knowledge through ontology. In this research, the ontology will be built automatically, which aims to reduce the role of human or expert knowledge to build ontology.

Related Work

An ontology is defined as an explicit specification of a conceptualization, which is an important tool for modeling, sharing and reuse of domain knowledge [2]. It allows domain knowledge to be represented explicitly through concepts and relations between them and hence to manipulate it automatically. However, ontology construction by hand is a complex and time consuming task [3]. Therefore, an automatic process is needed to help to facilitate the construction of ontology. Existing example approach to automated process is ontology design patterns (ODPs) [4].

The research of development of semi-automatic ontology using existing resources had been developed previously. Drame et al [5] builds a semi automatic-multilingual domain ontology is using UMLS Metathesaurus and parallel corpus. Validation of the ontology is constructed using Alzheimer's disease expert to ensure ontology constructed in accordance with the knowledge in Alzheimer's disease. However, this validation takes about a month to validate the ontology. It is take a lot of time. Therefore, this study developed using ODPs that validation can be done without the help of an expert. It aims to accelerate the development process ontology. The proposed method is an extension of conventional semi-automatic method.

Three studies related to the automatic ontology construction is research that conducted by Dahab et.al [6] who build automated construction ontology from natural language text. Then, Chen et.al [7] using recursive adaptive resonance Training (ART) network to construct a domain ontology-based TF-IDF. The study using web pages to build an ontology automatically. Navigli and Velardi [8] develop methodology for automatic ontology enrichment and document annotation.

Dahab et al build the domain ontology of natural text that using semantic pattern-based approach. This study analyzes the natural domain text to extract candidate relations and terms and mapping it into ontology. Meanwhile, Chen et.al using the Internet and web pages using HTML tags labels to choose the terms of web pages. Then calculated the TF-IDF to find weights of the used terms and then use the network ART (Adaptive Resonance Theory Network) to cluster terms. In study that conducted Navigli and Velardi, natural language definitions from available glossaries are processed and regular expressions are applied. The purpose is to identify general-purpose and domain-specific relations. The process in this research consist of pre-processing step (part-of-speech tagging and Named Entity Recognition), annotation of sentence segments with CIDOC properties and formalization of glosses. The evaluation methodology performance is extracting hypernymy and non-taxonomic relations. This study assessed the generality of the approach on a set of web pages from the domains of history and biography. The research in this paper is different from the three studies before because in this study

using a bilingual text corpora as a material to build ontology and using ontology approach design patterns (ODPs).

2. Methods

Resources

The BiblioDem Corpus
BiblioDem is a cumulative bibliographic database which currently contains 1556 scientific papers on Alzheimer's disease and related syndromes. This database contains abstracts of scientific papers selected from worldwide literature on Alzheimer's disease and their associated critical analysis, thus constituting rich knowledge. There are two kinds of corpus, namely corpus in English language and French language.

The corpus used in this research differs from previous research corpus [5]. Previous corpus contains scientific papers from 2004-2011, whereas I n this research using a corpus of scientific papers in 2013-2014 which amounts to 125 papers. The corpus can be obtained at the website address http://sites.isped.ubordeaux2.fr/bibliodem/bulletins.aspx.

Alzheimer Glossary
This research using Alzheimer glossary for filter extracted term from text corpora. The filter-ing term extraction is necessary because this process can produce the terms that have special relation with Alzheimer's disease, not term which is concerned with general health. Glossary Alzheimer can be obtained at the website address http://alzheimers.about.com/od/glossary/ and http://www.webmd.com/alzheimers/glossaryterms alzheimers. The total of terms which are related to Alzheimer disease in the Alzheimer glossary are 370 terms.

Ontology Design Patterns (ODPs)
Ontology Design Patterns (ODPs) can be accessed at http://www.gong.manchester.ac.uk/odp/html/index.html. This website also contains a catalog of ODPs. In this catalog, there are three types of ODPs namely (i) Domain Modelling ODPs, (ii) Good Practice ODPs (iii) Extension ODPs. The total number of ontology design patterns in the catalog number 16 ODPs. ODPs Domain Modeling aims to get the best model for a domain specific ontology. For example, Interactor_Role_Interaction and Sequence. Good Practice ODPs ontology aims to get better and stronger to maintain ontology models. For example, Normalization and Upper Level Ontology. On the other hand, ODPs Extension aims to overcome the limitations of existing ontology models to expand or increase coverage of the ontology. For example, Nary_DataType Relationship and Exception.

Tools

Text2Onto
Text2Onto is a framework of learning ontology which developed to support ontology construction from textual documents. Text2Onto has been used Cimiano and Volker [3]. The research used Text2Onto as a framework for ontology learning from textual resources based on Probabilistic Ontology Model (POM). There are three processes in Text2Onto: preprocessing, Execution of Algorithms and Combining results. During preprocessing, Text2Onto calls GATE application to tokenize document and tag Part of Speech sentences to creates indexes for the document and the result of this process is obtained as an annotation document. Execution of Algorithms is the process of Text2Onto executes the applied algorithms to extract terms and relations. One of the applied algorithm is TFIDF Concept Extraction. The last process is combining results, this process combines result of extracted terms and relations derived from processed documents. Text2Onto can be accessed at http://code.google.com/p/text2onto/downloads/list.

SimMetrics
SimMetrics is an open-source library available in Java which contains more than 20 similarity distance algorithms. For example, Jaro-Winkler, Levenstein distance, and Monge Elkan distance. SimMetrics used for string matching to identify the position of string or set of strings within a text. String matching algorithms helps to compare two different strings and look for similarity score between two text comparison. SimMetrics has been used by Chapman et al [9]. This research using simMetrics to calculate similarity between texts, where the information in this text will be integrated in a large repository (e.g. the Web). SimMetrics can be accessed at https://github.com/Simmetrics/simmetrics.

Ontology Generation
Ontology generation is a plugin in protégé to build ontology with generate terms of natural language text. Ontology generation was developed by Watcher and Schroeder, 2010 [10]. This tool supporting the creation and extension of OBO ontology by semi-automatically generating terms, definitions and parent–child relations from text in PubMed, the web and PDF repositories. This tool generates term by identifying significant noun phrases in text statistically and for the definitions and parent–child relations it employs pattern-

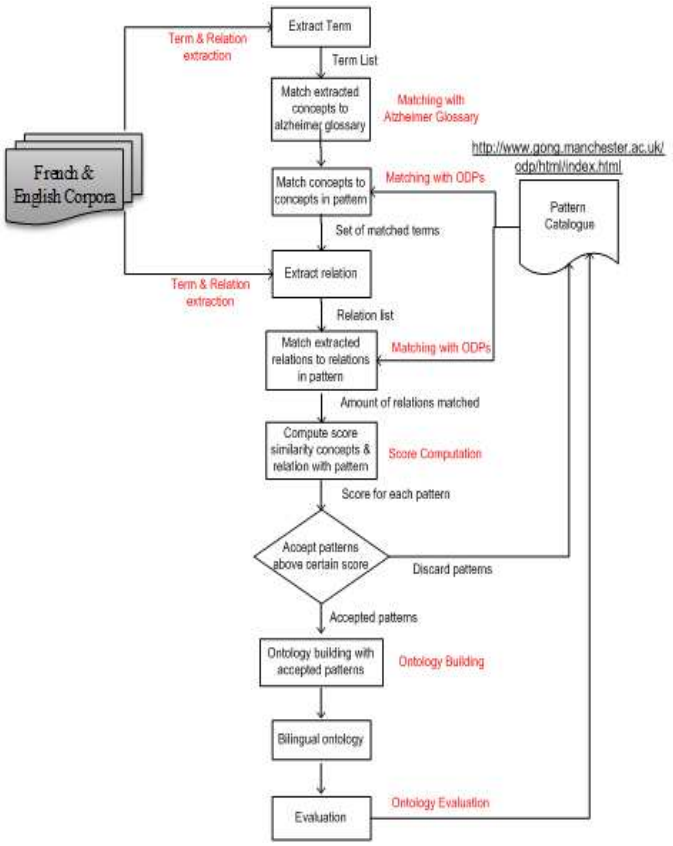

Figure 1. Overview of the ontology building method

based web searches. Ontology generation can be obtained at http://protegewiki.stanford.edu/wiki/Ontology_Generation_Plugin_(DOG4DAG). Ontology generation can be applied to the protégé-OWL version 4.1

Methods

The methods in this research consists of six stages: (i) Term and relation extraction (ii) Matching with Alzheimer glossary (iii) Matching the ontology design patterns (iv) Score computation similarity term and relations with ODPs (v) Ontology Building (vi) Ontology evaluation. To explain the process of each stage in the methodology is indicated in the figure 1.

The idea of the methodology is to take the extract-ed terms and relations, match them against the patterns and depending on the result use parts of the patterns to build the ontology. As a preprocessing step, a text corpus was analyzed by some term extraction software, which renders a list of possibly relevant terms. This list of terms is the input for this method.

Term & Relation Extraction

BiblioDem corpus extracted to obtain several terms. Further, relation which links between terms are extracted to obtain several relations. This corpus extraction using tools called Text2Onto.

Matching with Alzheimer Glossary

Having acquired several terms and relations, this terms and relation have matched with an Alzheimer glossary. At this stage, the matching with Alzheimer glossary is aimed to filter term so the same term derived from the extracted word list and list of words in the glossary.

Matching with ontology design patterns

The extracted terms and relations will be compared with terms and relations contained in Catalog ODPs that consist 16 design patterns. The matching result will be calculated score of similarity by SimMetrics tools that using Euclidean Distance algorithm. Then, two scores obtained from matched concepts and matched relations are weighted together to form a “total

TABLE 1
THE RESULT OF THE SIMILARITY CALCULATION ODPS

No	ODPs Type	Name	Similarity Value
1	Domain_Modeling ODP	Adapted_SEP	52%
2		Composite Property Chain	62%
3		Interactor Role Interaction	52%
4		List	46%
5		Sequence	51%
6	Extension ODP	Exception	42%
7		Nary Data Type Relationship	52%
8		Nary Relationship	54%
9	Good Practice	Closure	71%
10		Defined Class Description	56%
11		Entity Feature Value	47%
12		Entity Property Quality	56%
13		Entity Quality	55%
14		Normalization	50%
15		Selector	38%
16		Upper Level Ontology	17%
17		Value Partition	44%

matching-score" for each pattern. Then a decision is made according to some threshold value, the patterns will be kept and included in the ontology result, which will be discarded. Finally, an ontology is built from the accepted patterns that has the highest score similarity.

Score Computation

At this stage similarity calculation are computed between the extracted term and relation of the concepts and the relationships that exist in the design pattern. This stage using tools called SimMetrics. In this tool, there are various algorithms for example Euclidean Distance similarity distance, Levenshtein, and others. At this stage, average values are calculated from all the existing algorithms, so can be obtained value or score for string matching. The result in the score computation stage is the value or similarity score for each design pattern. Afterwards, a design pattern that has the highest similarity score is implemented to build ontology. We give more attention for relation between the concept because it can make ontology more structured.

Ontology Building

Ontology building is the stage to build an ontology of terms and relations that correspond to ontology design patterns. Ontology which implements a design pattern that has highest similarity value and Alzheimer ontology has been built on a previous study [5]. This stage uses tools to produce named OWL ontology generation to build ontology from terms and relationships that exist.

The step to using ontology generation is the first we must search definition of the term which entered. The search is connecting with PubMed in the protégé. After that, the automatic mapping of terms and relation that exist as to build a new ontology.

Ontology Evaluation

Ontology evaluation can be viewed in terms of complexity, time and effort required to build this ontology. This evaluation compared with the result by Drame et al [5] that construct semi-automatic ontology. Moreover, ontology evaluation also calculates accuracy of the terms and relation that used to build the ontology. Accuracy is calculated by equation(1).

$$\boldsymbol{accuracy = x/y} \quad (1)$$

Where, x is matching results of term/relation and y is total all of match term/relation.

The meaning of matching results of term or relation is match terms and relations that extracted from corpus with the terms and relations in design patterns that have the highest score similarity.

The meaning of total all term/relation is all of the terms and relations that extracted from corpus and has been filtered by Alzheimer's Glossary. That terms and relations are matched with the terms and relation on ODPs.

3. Results and Analysis

Term & Relation Extraction

The corpus used in this research includes 125 papers. The results of term and relation extraction are 1995 terms and 42 relations between terms. The number of terms resulting from the extraction of corpus was very large, so it needs to filtering terms that have association with Alzheimer's disease.

Matching with Glossary Alzheimer

At this stage, the result of term and relation extraction is the filtering with a matching Alzheimer's glossary. Alzheimer's glossary contains 370 terms related to Alzheimer. Once matched, the term acquired a number of 350 terms. This is different from the terms extracted from a corpus

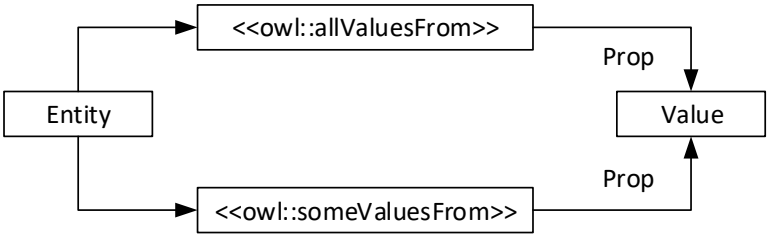

Figure 2. Structure of ODPs closure

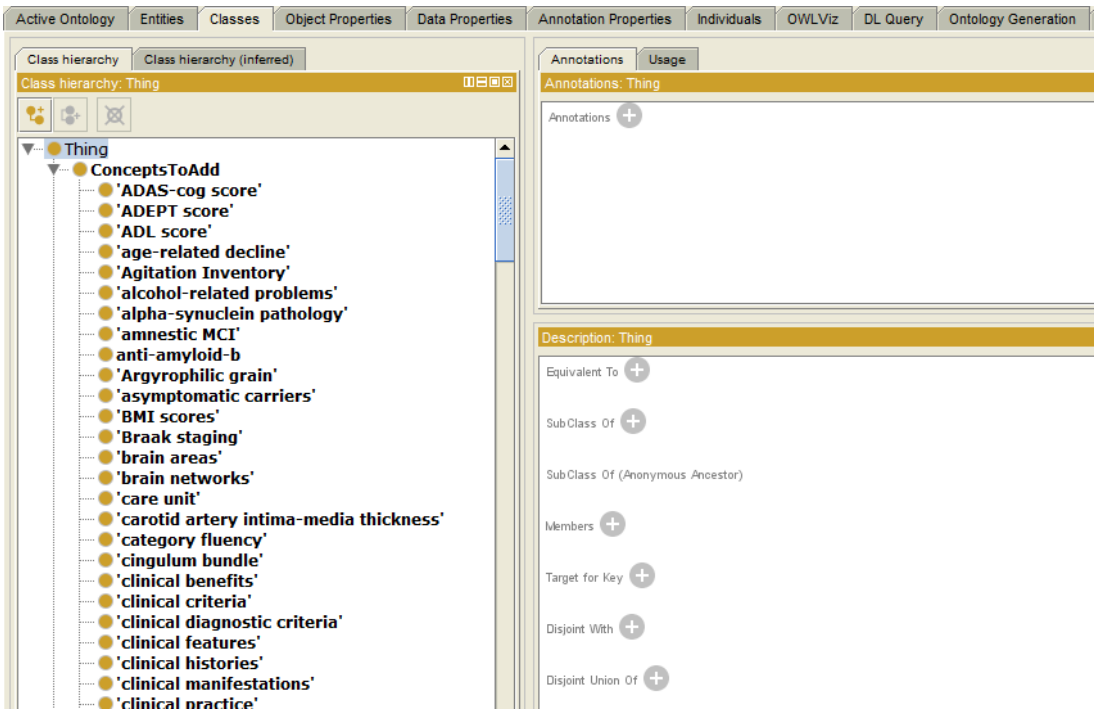

Figure 3. Visualization one part of the ontology in protégé

using extraction with text2onto because the extraction term will be related many health term in general, not specifically related to Alzheimer's disease. In addition, the number of terms in the Alzheimer's glossary of much less than terms of any health glossary in general so the scope of term filtering will be limited.

Matching with Ontology Design Patterns

Term and relation that has been filtered will be matched with a list of terms and relation that exist in the ontology design patterns. In the catalog there are several kinds of ontology design patterns (ODPs). The matching results will then be calculated for the similarity values between terms and filtering results with a term relation and relation that exist in the ontology design patterns (ODPs).

Score Computation

The result of similarity matching between term and relation with each ontology design patterns is shown in table 1.

The highest value of similarity found in ontology design patterns closure is equal to 71%. Closure ontology design pattern is a design pattern that limits the relationships among concepts which allows it to happen by clarifying the relation [11]. The limitations in this relation allows to express a concept has had a particular relation and only those relation. For example, a carnivorous is a meat eating animals, with closure design pattern can be revealed that carnivores do not eat other foods besides meat.

Ontology Building

Fully automatic ontology in this research consists of several components, there are 381 terms and

184 relations. Terms and relations are used to build ontology with tools namely OWL ontology generation. The figure 3 represents the results of the ontology that has been built in the protégé editor tool. There are 200 new terms and 42 new relations were added in that ontology.

The method of this ontology construction can be applied to other domains using a bilingual corpus associated with that domain and use the glossary or dictionary associated with that domain to filtering the terms and relations associated with that domain. If there is a bilingual corpus and glossary related to a specific domain ontology that can be built using the stage as a method of this research

Ontology Evaluation

Ontology evaluation can be viewed in terms of complexity, time and effort required to build this ontology. The result of evaluation is fully automatic ontology construction that can shorten the development time compared to ontology manually or semi-automatic which requires expert validation for a month. In previous studies it takes two teams in the field of Alzheimer's expert to validate the built of ontology. New term and relation in fully automatic ontology construction present that the ontology more complexity than semi automatic ontology in previous research [5].

The result of accuracy value of fully automatic ontology construction is 72%. It is obtained from the calculation of the number of terms or relations corresponding number of 525 terms or relations and the total term or relation in the ontology built a number of 726 terms or relations. This indicates that fully automatic ontology construction method used in the study was quite nice to be able to build the ontology, but it still needs to be improved in order to obtain higher accuracy values.

This accuracy value can not be general in this research, because the accuracy value can be different for other cases. However, the accuracy value can be as a supporting material to the evaluation of this research.

4. Conclusion

This research succeeds to make fully automatic bilingual domain ontology using the Ontology Design Patterns (ODPs) and text corpora. The result of ontology development includes 381 terms and 184 relations with addition of 200 terms and 42 new relations.

Fully automatic construction could speed up and reduce the human's role as expert to evaluate ontology rather than building ontology manually. The result of evaluation is fully automatic ontology construction that can shorten development time compared to manual ontology or semi-automatic which requires expert validation. New term and relation in automatic ontology construction present that the ontology are more complicated than semi automatic ontology in previous research.

For future work, addition number of term in Alzheimer's glossary is recommended to filter the term results of a corpus extraction well. Alzheimer's glossary can improve the results of filter term from extraction corpus. In addition, type of data ontology design patterns (ODPs) can be improved to get the highest similarity value for selected design patterns that will be implemented to build ontology.

Moreover, ontology enrichment to increase the number of terms can be implemented in ontology building. Ontology enrichment using parallel corpora of the website in English and French can obtain terms and synonymous terms in other languages. This process uses term alignment with alignment approach to enrich bilingual biomedical resource. After that, parallel term can integrate into ontology that has been built.

Fully automatic ontology construction not only can be applied to Alzheimer's domain knowledge, but also fully automatic ontology construction methods can be applied to other domain knowledge. Therefore, ontology construction on other domain ontology can be used for research and development of their domain knowledge.

References

[1] Louis, Jean L. Prototype System For Automatic Ontology Construction. Thesis Magister Information Technology. The Royal Institute Of Technology. Sweden. 2007.

[2] Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisit 1993;5:199–220.

[3] J. Cimiano, Philipp and Völker, "A Framework for Ontology Learning and Data-Driven Change Discovery," Nat. Lang. Process. Inf. Syst., pp. 227– 238, 2005.

[4] Blomqvist, E. Fully Automatic Construction of Enterprise Ontologies Using Design Patterns: Initial Method and First Experiences. In Proceedings of OTM 2005 Conferences, Ontologies, DataBases, and Applications of Semantics (ODBASE), Agia Napa, Cyprus, Oct 31- Nov 4, 2005.

[5] Dramé K et al. Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: An application

to Alzheimer's disease. J Biomed. Inform., vol. 48, pp. 171–182, 2014.

[6] Dahab, M.Y., Hassan, H. and Rafea, A., "TextOntoEx: Automatic ontology construction from natural English text," Expert Syst. Appl., vol. 34, pp. 1474–1480, 2008.

[7] Chen, R. C., Liang, J. Y., and Pan, R. H., "Using recursive ART network to construction domain ontology based on term frequency and inverse document frequency," Expert Syst. Appl., vol. 34, pp. 488–501, 2008.

[8] Navigli, R., and elardi, P., "From Glossaries to Ontologies : Extracting Semantic Structure from Textual Definitions," Ontol. Learn. Popul. Bridg. Gap between Text Knowl., pp. 71–87, 2008.

[9] Chapman, S., Norton, B., and Ciravegna, F., "Armadillo: Integrating knowledge for the semantic web," Proc. Dagstuhl Semin. Mach. Learn. Semant. Web, pp. 2–4, 2005.

[10] Wächter, T., and Schroeder, M., "Semi-automated ontology generation within OBO-Edit," Bioinformatics, vol. 26, pp. 88–96, 2010.

[11] ODP public catalog. Closure. http://www.gong.manchester.ac.uk/odp/html/Closure html. Access on Monday, 26 May 2014. 09.00 am.