IPASJ International Journal of Information Technology (IIJIT)
Web Site: http://www.ipasj.org/IIJIT/IIJIT.htm
A Publisher for Research Motivation ..... Email:editoriijit@ipasj.org
Volume 5, Issue 2, February 2017 ISSN 2321-5976

POPULATING DOMAIN EXACT WORDS FROM INSTRUCTIVE SITES OF STATE UNIVERSITIES TO CREATE DOMAIN METAPHYSICS FOR ACADEMIC WEBSITES

Mr. Dhrendra Sinna

Sagar Institute of Science and Technology, Bhopal

ABSTRACT

AI from one linguistic communication to the opposite may be a difficult task. one amongst the strategies of doing AI is mistreatment artificial language primarily based approach. in this approach the language may be delineated in associate degree intermediate type, which may be translated to the target language. Generation of linguistic communication sentence combines data concerning language and also the application domain to supply correct translation. And thus, it's vital to arrange domain-specific corpus. conjointly it's equally vital that the linguistics hierarchy among the sets of domain words for AI of a document, since the hierarchy can offer linguistics links and metaphysics info for words. Ontologies outline ideas and interrelationships so as to supply a shared vision of a given application domain. one amongst the most issues is that the issue in distinguishing and process relevant ideas within the domain. This paper aimed the extraction of data from state university websites, so as to spot the domain specific words for academic sites. This paper proposes a technique to spot domain specific words by utilizing the data structure of net directories node-by-node. This technique can turn out an inventory of domain dependent words with high frequency words.

1.INTRODUCTION

linguistic communication process is presently an energetic analysis space. this can be as a result of most of {the net|the online|the net} sites or in different words the knowledge within the web is in English however the non-English language users of the net square measure on the rise in per annum. and so they're to be delivered within the type of their linguistic communication. linguistic communication understanding may be delineated because the conversion of linguistic communication into a laptop processable data illustration that ultimately conveys the linguistics interpretation of the text. the particular illustration will vary from easy extracted keywords to advanced logical, graph or frame like structures ([1] Brachman & Levesque 1985). The goal of the linguistic communication process cluster is to style and build software package that may analyze, understand, and generate languages, that square measure handled naturally by humans. a number of the basic tasks related to linguistic communication process embody morphological analysis, elements of Speech (POS) Tagging, Named Entity Recognition, Multiword Expression Extraction, Shallow Parsing, linguistics Interpretation and at last Pragmatic and Discourse process ([2] Bhattacharyya 2012). Natural languages square measure inherently ambiguous within the sense that a word, phrase or a sentence will have totally different interpretations ([3] Udemmadu 2012). In fact, the varied sorts of ambiguity (e.g. lexical, semantic, denotive, etc.) square measure one amongst the most challenges for try linguistic communication process. resolution those ambiguities is critical to create refined systems dedicated to applications like Question respondent, AI, info retrieval etc. one amongst the foremost vital tasks of linguistic communication process is developing a full-fledged bilingual AI system for any 2 natural languages that may be a difficult and demanding task ([4] Antonius P.J, 2013). Lexical resources or knowledge domain play a very important role in linguistic communication process tasks particularly within the case of AI. Not solely the lexicons, however the linguistics hierarchy of the lexicons are vital for AI. Since the stratified sites offer linguistics tag info (explicitly type the HTML/XML tags or implicitly from the directory names) and helpful linguistics links, it's fascinating that the lexicon construction can be conducted mistreatment the net corpora.

In order to exchange info across cultures and languages, it's essential to form design to share varied lexical resources across languages. Universal Networking Language is associate degree design that is employed to represent the languages for the aim of AI. UNL illustration is that the artificial language structure used for representing the linguistics of the languages. UNL is semantically biased language freelance. Universal Word (UW), outlined by a

headword associate degreed a group of restrictions that provide an unambiguous illustration of the construct, forms the vocabulary of Universal Networking Language. There exist lexical gaps not solely as a result of that a word in one language has no correspondence in another, however there square measure variations within the ways that languages structure their words and ideas ([5]Pease and Fellbaum 2010). Domain metaphysics reduces or eliminates the abstract and word confusion among the members of virtual community of users (for example, traveler operators, man of science, students, industrial enterprises) that require to share electronic documents and knowledge of assorted sorts. this can be achieved by distinguishing and properly outlined set of relevant ideas that characterise a given application domain. associate degree Ontoloy is thus a shared understanding of some domain of interest. tho' there square measure several higher domain metaphysics, the provision of Specific domain ontologies that square measure essential to beat the barrier of actual inconsistencies. General purpose resources like WordNet ([6] Niles and Pease, 2003). et al. deals with thousands of ideas, they are doing not encrypt abundant of the domain data required by specialised applications. Although domain ontologies are recognized as crucial resources for translation, in practice, full-fledged resources are not available. The purpose of this study is to aid translation of academic web sites in English to Tamil language. For which the domain ontology for academic web sites needs to populated and represent them as UNL Ontology. Towards that aim the first step is to identify the domain dependent words. This paper presents a methodology to extract and build domain-specific corpora and represent tem in the form of ontology for educational sites. Different steps are necessary for such task. Section 2 presents the Universal Networking Language used to represent the Interlingua and ontology used to represent domainspecific corpora. Section 3 specifies the methodology; section 4 concludes and hints at future work.

2. OVERVIEW OF UNL

Semantic Relation aims at giving a semantic relationship that exists between any two concepts in a sentence. Semantic relations are unidirectional underlying connections between concepts. Semantic relations are the building blocks for creating the semantic structure of a sentence. There are four different types of semantic relations listed by Grabar & Hamon (2004). They are namely Lexical (Synonymy), vertical(hypernymy, meronymy) and domain-specific relations. One of the many representations of generic semantic relation is the UNL representation. Any translation system using UNL as intermediate representation needs to have an EnConverter from the source language to UNL and a DeConvereter from UNL to target language. Universal Networking language is an electronic language in the form of semantic network that act as an intermediate representation to express and exchange every kind of information. This language is assumed to express meanings in the same standardized way as HTML represents its layout. The UNL represents information sentence by sentence. Sentence is represented as a hyper-graph having universal Words (UWs) as nodes and relations as arcs. This hyper-graph is also represented by a set of directed binary relations between two of the UWs present in the sentence. Nodes or Universal Words are words based on English and disambiguated by their positioning in a Knowledge Base (KB) of conceptual hierarchies. The text once converted into UNL can be converted to many different languages, for instance, once a home page is expressed in UNL, it can be read in a variety of natural languages. Furthermore, if the type of knowledge required for doing some task is described in a language, such as UNL, the software only needs to interpret unambiguous intermediate instructions written in the language to be able to perform its function. As a result of this standardized meaning representation, documents no longer need to be multiplied in order to represent the content in different natural languages. The meaning representation is directly available to retrieval and indexing mechanisms and tools for automatic summarizing and knowledge extraction, and it will be converted to a natural language only when communicating with a human user. The task of representation of a UNL web-page to a web user will be taken over by a UNL-Viewer. In ne commercially homeward state of affairs, the UNL-viewer represents a replacement generation of web-browser that additionally to their capabilities to handle java and java-script, square measure equipped with one or additional national UNLDeconverter so as to show the which means content in a very national language.

A. Universal Words

Universal Words square measure words of the UNL that represents the UNL vocabulary. they're the labels for ideas, syntactic-semantic units that mix to create UNL expressions. each UW denotes a thought. The which means of a sentence is expressed by the mixture of a group of UWs that square measure joined by relations and changed attributes. A UNL illustration may be a hyper-graph within which the UWs square measure nodes, or arguments of the binary relations. each UW ought to be outlined within the UNL knowledge domain. A UW itself doesn't itself convey its entire which means. A UW is taken by relating all its doable relations with different UWs. These relations square measure outlined within the UNL computer memory unit, so as to render a UW substantive, by making links with these relations within the UNL computer memory unit.

B. UNL Relations

Binary relations square measure the building blocks of UNL sentence. they're created from a relation and 2 UWs. Relations that link UWs square measure tagged with linguistics roles of the sort like agent, object, experience, time, place, cause, that characterise the relationships between the ideas collaborating within the events or states a natural sentence. UNL has mere forty such relations and claim that these relations square measure comfortable to represent the interconnection expressed by linguistic communication sentences.

D. UNL metaphysics

In The goal of domain metaphysics is to cut back or eliminate the abstract and word confusion. this can be achieved by distinguishing and properly process a group of relevant ideas that characterise a given application domain. Ontologies could have totally different degrees of ritual however they essentially embody a vocabulary of terms with their which means and their relationships.

3. METHODOLOGY

Domain-specific corpora in languages aside from English aren't as simply found. Language translation so desires domain-specific corpora, however the matter reside within the overhead work concerned in building such corpus. It involve the

- method of choice of sources
- extracting the domain-specific words
- representing them in associate degree accessible type

Since the net documents nearly type a particularly Brobdingnagian document classification tree, it's planned to convert it into a lexicon tree, and assign implicit tags to the domain specific words within the net document mechanically. This approach is impressed by the very fact that almost all sites within the websites square measure already classified in a very stratified manner; the stratified directory structures implicitly counsel that the domain specific terms within the text materials of a specific directory square measure closely associated with a typical subject, that is known by the name of the directory. If it's detected a website specific words among every document, and take away words that square measure non-specific and tag the DSW's so noninheritable with the directory name, then it's doable to nearly get a stratified lexicon tree. In such a tree, every node is semantically joined by the initial net document hierarchy, and every node encompasses a set of domain specific words related to it.

In the extraction method, the directory names may be thought to be implicit sense label or implicit linguistics tags, and also the action to place the net pages into properly named directories may be thought to be implicit aging method by the net masters. and also the hierarchy itself provides info on the hierarchy of linguistics tags. From a well-organized computing device, it's doable to amass associate degree implicitly labelled corpus from that web site. And so there's no price to extract DSW's from such net corpora. This paper thus uses the tactic of constructing lexicon-tree from the net hierarchy, wherever domain specific word identification seems to be a key issue and also the start towards the development method. Since the terms (words or compound words) within the documents embody general terms further as domainspecific terms, the sole downside is a good model to execute those domain-independent terms from the implicit tagging method. The degree of domain freedom may be measured with the inter-domain entropy. Generally, a term that distributes equally altogether domains is probably going to be freelance of any domain. thus such terms may be weighted less for it to be a probable Domain Specific word.

2. Once the list of URLs for every page has been found it may be placed within the tool TextSTAT-2.9 to gather keywords. Collocations and concordance found within the terms offer valuable info for effort the sense and usage of a term or word. Concordances square measure typically outlined clearly as a window of text encompassing a term or expression of interest. Most often, a set tiny window size is established and also the results square measure referred to as Keyword in context (KWIC). Collocations square measure words that tend to go with on top of a random chance. though conceptually the definition is kind of easy, results can mostly dissent due to 2 main variables. the primary variable is that the window size within which co-occurrences square measure measured. atiny low window is typically established for collocations.

4. EVALUATION

All the colleges of state has been collected from net. so as to indentify Domainindependence words different sites like New sites were used. Table one shows a number of the domain-specific words extracted. as an example the word "faculty" is specially utilized in the tutorial internet sites, wherever as "news" is employed in broadcast domain, and

thus, the domain specific words and their domain tags square measure annex associated. As a results of such association, low inter-domain entropy words within the same domain are extremely correlative. It can even notice lexicon relations among domain tags and domain specific words from table one.

5. CONCLUSION AND FUTURE WORK

The current state of affairs there's no UNL lexicon for Tamil language fully type. we have a tendency to tried to form a vocabularies associated with academe. The metaphysics for a similar also will be created for proper translation in to Tamil. For the nonce we've focused solely state Universities internet sites. The domain specific words of educational sites are inhabited. we've achieved one hundred fifty words and that they need to be organized within the hierarchy of UNL metaphysics.

6. REALATED WORK

Emhimed Salem Alatrish et. al. have planned a semi-automatic procedure to form ontologies for various natural languages[8]. It aims to integrate totally different software package tools that provides the building of ontologies for various natural languages. Nitsan chrizman et. al. have centered o automatic construction of multi-lingual domain-ontologies[9]. during this work it aims to form DAG(Directed Acyclic Graph) that consists of the ideas associated with a particular domain and also the relations between them. Hadhemi Achour et. al. have planned linguistics metaphysics primarily based model for polyglot categorisation and retrieving the tutorial resources of an internet learning environment[10]. It uses a image to perform multilingual looking out (Arabic-English-French) for on-line learning resources. It uses categorisation a info of learning resources associated with the theme of object homeward programming in Java of the domain of technology.

REFERENCES

[1]. Levesque, Hector; Ronald Brachman (1985). "A elementary exchange in data illustration and Reasoning". In Ronald Brachman and Hector J. Levesque. Reading in data illustration. Morgan Kaufmann. p. 49. ISBN 0-934613-01-X
[2]. Pushpak Bhattacharyya and Subhabrata Mukherjeet, sentiment analysis in tiwtter with lightweight discourse analysis, Proceedings of COLING 2012, Techniczl pape, pages 1847-1864, Mumbai, December 2012.
[3]. Thecla-Obiora Udemmadu,The Issue of Ambiguity in the Igbo Language , AFRREV LALIGENS An international journal of language, literature and gender studies, Vol 1(1) March, 2012:109-123.
[4]. Antony P.J., Machine Translation Approaches and survey for Indian Languages, International journal of computational linguistics and Chinese language processing, Vol 18, no. 1, pp. 47-78.
[5]. Adam Pease and Christiane Fellbaum. 2010. Formal ontology as interlingua: The SUMO and WordNet linking project and global wordnet. In Ontology and Lexicon, A Natural Lnaguage Processing Perspective, pages 25-35. Cambridge University Press.
[6]. Ian Niles and Adam Pease. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the suggested Upper Merged Ontology. In proceedings of the 2003 International Conference on Information and Knowledge Engineering(IKE 03), Las Vegas, Pages 412-416.
[7]. Khan Md. Anwarus Salam, Hiroshi Uchida, Setsuo Yamada, Tetsuro Nishino, net primarily based UNL metaphysics visual image, Journal of convergence info technology, volume 8, number 13, August 2013, pp:69-75.
[8]. Emhimed Salem Alatrish, Dusan Tosic and Nikola Milenkovic,"Building ontologies for various natural languages", technology and knowledge Systems 11(2):623-644, DOI:10.2298/CSISI30429023A, 2014
[9]. Nitsan Chrizman and Alon Itai, "How to construct polyglot Domain Ontologies",[online]. URL:www.irecconf.org/proceedings/Irec2014/pdf.730_paper.pdf, 2014.
[10]. Hendez, M., Achour, H. "Keywords extraction for automatic categorisation of e-learning resources", laptop Applications & analysis (WSCAR), World conference on pp: 1-5, 2014.