Populating domain specific words from Academic web pages of Tamil Nadu Universities to build domain ontology for educational websites.

Dr. K.Ponmozhi

Department of Information Technology
Hajee Karutha Rowther Howdia College
Uthamapalayam,Theni, India.

Abstract

Machine translation from one natural language to the other is a challenging task. One of the methods of doing machine translation is using Interlingua based approach. In that approach the source language can be represented in an intermediate form, and that can be translated to the target language. Generation of Natural language sentence combines knowledge about language and the application domain to produce correct translation. And thus, it is important to prepare domain-specific corpus. Also it is equally important that the semantic hierarchy among the sets of domain words for machine translation of a document, since the hierarchy will provide semantic links and ontological information for words. Ontologies define concepts and interrelationships in order to provide a shared vision of a given application domain. One of the main problems is the difficulty in identifying and defining relevant concepts in the domain. This paper aimed the extraction of knowledge from Tamil Nadu university websites, in order to identify the domain specific words for educational sites. This paper proposes a method to identify domain specific words by utilizing the hierarchical structure of web directories node-by-node. This method will produce a list of domain dependent words with high frequency words.

Keywords- Machine Translation; UNL; Natural Language Processing; Domain knowledge.

Introduction

Natural Language Processing is currently an active research area. This is because most of the web sites or in other words the information in the web is in English but the non-English language users of the web are on the increase in every year. And therefore they are to be delivered in the form of their native language. Natural language understanding can be described as the conversion of natural language into a computer processable knowledge representation that ultimately conveys the semantic interpretation of the text. The actual representation can vary from simple extracted keywords to complex logical, graph or frame like structures ([1] Brachman & Levesque 1985).

The goal of the Natural Language Processing group is to design and build software that will analyze, understand, and generate languages, which are handled naturally by humans. Some of the fundamental tasks associated with Natural Language Processing include morphological analysis, Parts of Speech (POS) Tagging, Named Entity Recognition, Multiword Expression Extraction, Shallow Parsing, Semantic Interpretation and finally Pragmatic and Discourse Processing ([2] Bhattacharyya 2012).

Natural languages are inherently ambiguous in the sense that a word, phrase or a sentence can have different interpretations ([3] Udemmadu 2012). In fact, the various kinds of ambiguity (e.g. lexical, semantic, referential, etc.) are one of the main challenges for tackling Natural Language Processing. Solving those ambiguities is necessary to build sophisticated systems devoted to applications like Question answering, Machine Translation, Information retrieval etc.

One of the most important tasks of Natural language processing is developing a full-fledged bilingual Machine translation system for any two natural languages which is a challenging and demanding task ([4] Antony P.J, 2013). Lexical resources or knowledge base play an important role in natural language processing tasks especially in the case of machine translation.

Not only the lexicons, but the semantic hierarchy of the lexicons are also important for machine translation. Since the hierarchical web pages provide semantic tag information (explicitly form the HTML/XML tags or implicitly from the directory names) and useful semantic links, it is desirable that the lexicon construction could be conducted using the web corpora.

In order to exchange information across cultures and languages, it is essential to create architecture to share various lexical resources across languages. Universal Networking Language is an architecture which is used to represent the languages for the purpose of machine translation.

UNL representation is the Interlingua structure used for representing the semantics of the languages. UNL is semantically biased language independent. Universal Word (UW), defined by a headword and a set of restrictions which give an unambiguous representation of the concept, forms the vocabulary of Universal Networking Language. There exist lexical gaps not only because that a word in one language has no correspondence in another, but there are differences in the ways languages structure their words and concepts ([5]Pease and Fellbaum 2010).

Domain Ontology reduces or eliminates the conceptual and terminological confusion among the members of virtual community of users (for example, tourist operators, computer scientist, students, commercial enterprises) that need to share electronic documents and information of various kinds. This is achieved by identifying and properly defined set of relevant concepts that characterise a given application domain. An Ontoloy is therefore a shared understanding of some domain of interest.

Though there are many Upper domain ontology, the availability of Specific domain ontologies which are essential to overcome the barrier of actual inconsistencies. General purpose resources like WordNet ([6] Niles and Pease, 2003). and others deals with thousands of concepts, they do not encode much of the domain knowledge needed by specialized 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 domain-specific corpora. Section 3 specifies the methodology; section 4 concludes and hints at future work.

I. 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 oriented scenario, the UNL-viewer represents a new generation of web-browser which in addition to their capabilities to handle java and java-script, are equipped with one or more national UNL-Deconverter in order to display the meaning content in a national language.

The UNL-documents to be made available in the Internet are prepared neither manually nor fully automatically. The formal and linguistic specifications of this language are far too complex to be fulfilled by an untrained and unsupported person. Therefore, the creation of UNL-documents is supported by national

Enconverteres which convert a natural language text in a raw version of UNL. This raw version is to be visualized and edited in UNL-editors, a tool currently developed by the UNL-network to be used by trained UNL writers to finalize the UNL document.

A. Universal Words

Universal Words are words of the UNL which represents the UNL vocabulary. They are the labels for concepts, syntactic-semantic units that combine to form UNL expressions. Every UW denotes a concept. The meaning of a sentence is expressed by the combination of a set of UWs which are linked by relations and modified attributes. A UNL representation is a hyper-graph in which the UWs are nodes, or arguments of the binary relations.

Every UW should be defined in the UNL Knowledge base. A UW itself does not itself convey its entire meaning. A UW is interpreted by referring to all its possible relations with other UWs. These relations are defined in the UNL KB, in order to render a UW meaningful, by creating links with these relations in the UNL KB.

B. UNL Relations

Binary relations are the building blocks of UNL sentence. They are made up of a relation and two UWs. Relations that link UWs are labelled with semantic roles of the type such as agent, object, experience, time, place, cause, which characterise the relationships between the concepts participating in the events or states a natural sentence. UNL has specified forty such relations and claim that these relations are sufficient to represent the interconnection expressed by natural language sentences.

In addition to propositional content("who did what to whom"), the UNL expressions are intended to capture pragmatic information such as focus, reference, speaker's attitudes and intentions, speech, act and other type of information. This information is rendered by means of attributes attached to the nodes. Some of the UNL relations are:

agt - agent(amuthan runs),

aoj- a thing which is in a certain state or is ascribed a property (my brother is a student)

dur – duration (we worked eight hours)

fmt – a range between two things (we worked from monday till friday)

pos – possession(trust's institution).

These relations are classified into different categories. They are the Participant, Change of state, place, circumstantial, logical, conditional, temporal and numerical relations. There are many factors to be considered in choosing an inventory of relations. The principles to choose relations are as follows:

- Necessary Condition: When a UW has relations with more than two other UWs each relation label should be set in order to indentify each relation on the premise.
- Sufficient condition : When there are relations between UWs, each relation label should be set in such a way, as to understand each role of each UW only by refereeing to a relation label.

In order to handle the wide range of subjects which can be found in the Internet, the integration of a statistical component into the transfer component become necessary. This statistical component helps to select the ND which is most likely to be used in a given subject domain (e.g. 'medicine', 'sports', 'computing' 'education' etc.). For this purpose, the source language lexicon is supplied with the frequency information concerning English words in different subject domains.

C. UNL Knowledge Base

The UNL Knowledge Base is a semantic network comprising every directed binary relation between UWs. Every UW must be defined in the UNL KB and must also be linked to other related UWs. In UNL KB, all UWs are linked to each other through "icl","iof","equ" relations. The UW system forms a hierarchy of UWs with the use of these relations.

D. UNL Ontology

In The goal of domain ontology is to reduce or eliminate the conceptual and terminological confusion. This is achieved by identifying and properly defining a set of relevant concepts that characterise a given application domain. Ontologies may have different degrees of formality but they necessarily include a vocabulary of terms with their meaning and their relationships. Thus construction of ontology requires a thorough domain analysis that is accomplished by

- Carefully identifying the vocabulary that is used to describe the relevant concepts within the domain
- Coding complete and rigorous definition about the terms(concepts) in the vocabulary
- Characterizing the conceptual relations among those terms.

The semantics hierarchy among words (especially among sets of domain specific words) as well as the membership of domain specific words are important because the hierarchy will provide semantic links and ontological information such as is-an-instance-of and is-a-kind-of relationships among words.

UWs, the basic concepts of UNL are placed in a UNL Ontology ([7] Khan et al 2011). The UNL Ontology is a taxonomy of relations and is a tree-like structure where UWs are interconnected through hierarchical relations such as "icl" (is-a-kind-of) and "iof"(is-an-instance-of). The UWs in the UNL Ontology are divided into four major categories. They are:

1. Adverbial concepts which describe the manner (icl>how)
2. Attributive concepts which describes the modifies such as mod<thing and qua<thing
3. Nominal concepts which describe the abstract, attributive, concrete, functional and volitional things in addition to the concepts representing place and time (icl>thing)
4. Predictive concepts which describe the verbal concepts such as be (aoj>thing), do(agt>thing) and occur(obj>thing).

II. METHODOLOGY

Domain-specific corpora in languages other than English are not as easily found. Language translation indeed needs domain-specific corpora, But the problem reside in the overhead work involved in building such corpus. It involve the

- *process of selection* of sources
- *extracting the domain-specific words*
- *representing them in an accessible form*

Since the web documents virtually form an extremely huge document classification tree, it is proposed to convert it into a lexicon tree, and assign implicit tags to the domain specific words in the web document automatically.

This approach is inspired by the fact that most web pages in the websites are already classified in a hierarchical manner; the hierarchical directory structures implicitly suggest that the domain specific terms in the text materials of a particular subdirectory are closely related to a common subject, which is identified by the name of the subdirectory.

If it is detected a domain specific words within each document, and remove words that are non-specific and tag the DSW's thus acquired with the directory name, then it is possible to virtually get a hierarchical lexicon tree. In such a tree, each node is semantically linked by the original web document hierarchy, and each node has a set of domain specific words associated with it.

For instance, a subdirectory entitled 'examination' is likely to have a web pages containing domain specific terms 'courses', 'fees', 'timetable' and so on. And thus it is possible to collect the domain specific words from such a directory.

In the extraction process, the directory names can be regarded as implicit sense label or implicit semantic tags, and the action to put the web pages into properly named directories can be regarded as implicit aging process by the web masters. And the hierarchy itself provides information on the hierarchy of semantic tags.

From a well-organized web site, it is possible to acquire an implicitly tagged corpus from that site. And thus there is no cost to extract DSW's from such web corpora.

This paper therefore uses the method of constructing lexicon-tree from the web hierarchy, where domain specific word identification turns out to be a key issue and the first step towards the construction process.

Since the terms (words or compound words) in the documents include general terms as well as domain-specific terms, the only problem is an effective model to execute those domain-independent terms from the implicit tagging process. The degree of domain independency can be measured with the inter-domain entropy. Generally, a term that distributes evenly in all domains is likely to be independent of any domain. Therefore such terms can be weighted less for it to be a probable Domain Specific word.

Domain specific word extraction algorithm:

1. Acquire a large collection of web documents using a web spider while preserving the directory hierarchy of documents. Strip unused mark up tags from the web pages.

This paper used web sites of Tamil Nadu Universities. The page links of those sites can be got from the tool Visual Web Spider. Visual Web Spider is a web crawler tool used in this work to find the page URLs of given the web sites intended. It provides various options to specify filters, encoding, exporting methodologies. We can specify crawling rules such as whether to follow all internal links, external links, whether to allow URL if they contain .jpg,.bmp,google.com etc. We can also specify maximum crawling depth. We have used 10. The

data can be exported to MS Access database or MySQL Database or can be as Excel, HTML files. Figure 1 specifies the screen shot of alagappa university.

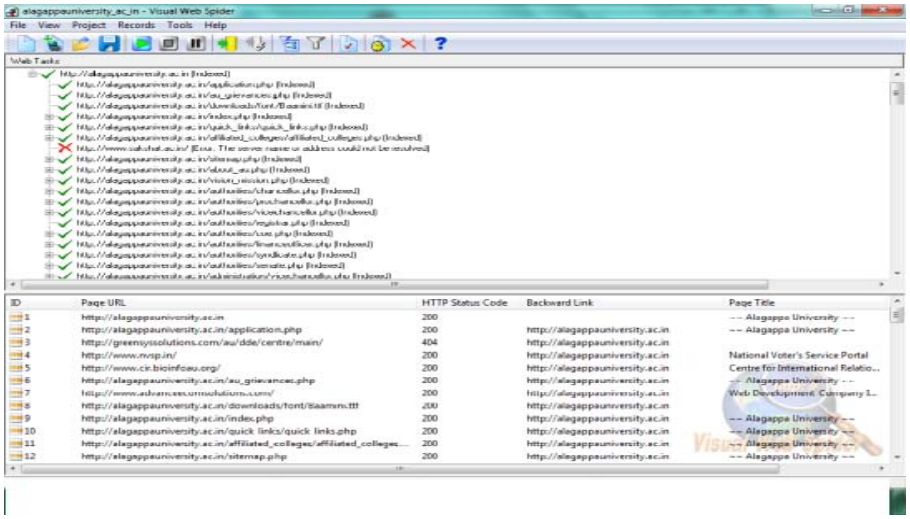

Figure 1. Visual web spider for Alagappa university

2. Once the list of URLs for each page has been found it can be placed in the tool TextSTAT-2.9 to collect keywords.

Collocations and concordance found in the terms provide valuable information for acquiring the sense and usage of a term or word.

Concordances are usually defined clearly as a window of text surrounding a term or expression of interest. Most often, a fixed small window size is established and the results are called Keyword in context (KWIC). Collocations are words which tend to co-occur with higher than a random probability. Although conceptually the definition is quite simple, results will largely differ because of two main variables. The first variable is the window size in which co-occurrences are measured. A small window is usually established for collocations. A second variable is the actual measure of association used, and there have been multiple measures suggested in the literature, such as Overlap, Manual Information, Dice Coefficient etc.

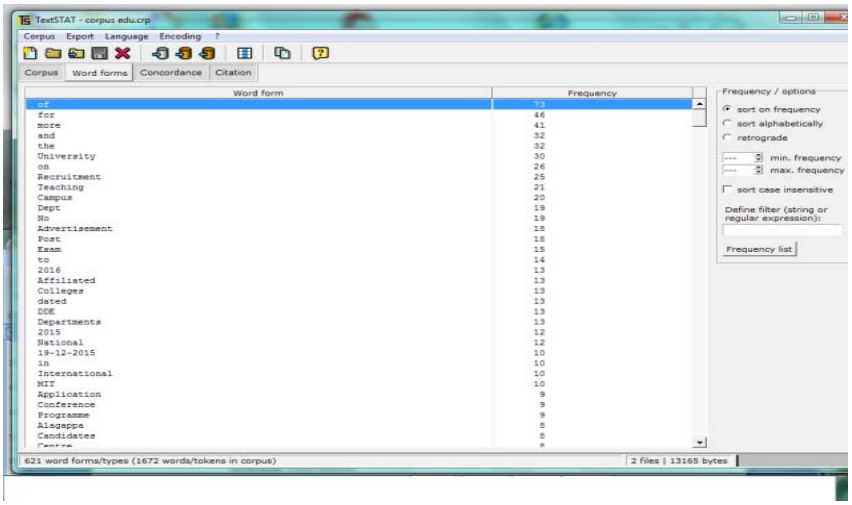

Figure 2: Screen shot of word frequency list of Alagappa university home page

3. The Acquiring Normalized Term Frequencies for all Words in Various Domains, by assigning a set {Wi,Dj,fij} ,where wi is the ith words of the complete word list for all documents , dj is the jth directory name, and fij is the normalized relative frequency of occurrence of wi in the directory dj.
4. Identifying the domain independent terms which are distributed evenly in all directories. It is therefore assigned a less weight to these terms. The high frequency terms may be more important in a domain. These terms are assigned with high weight.
5. The words are displayed with decreasing order of weightage values.

III. EVALUATION

All the Universities of Tamil Nadu has been collected from internet. In order to indentify Domain-independence words other sites such as New sites were used.

Table 1 shows some of the domain-specific words extracted. For instance the word "faculty" is specially used in the academic web sites, where as "news" is used in broadcast domain, and thus, the domain specific words and their domain tags are ell associated.

As a result of such association, low inter-domain entropy words in the same domain are also highly correlated. It can also find lexicon relations among domain tags and domain specific words from table 1.

Hypernym/Hyponym: University vs College ; News vs District news

Has-Member/ Member Of: faculty vs teaching staff, non-teaching staff

Has-Part/Part-of: facilities vs WiFi

Antonym: Pass vs Fail

Such lexical relations are, in general interested to build Domain specific Ontologies. Extracting DSW's with the inter-domain entropy metric is therefore well founded.

TABLE I. FRQUENTLY FOUND WORDS IN DOMAINS OF ACADEMIA AND NEWS SITES

Academic	News
University	Current news
Department	Horoscope
Conference	Sports
Examination	Editors note

IV. CONCLUSION AND FUTURE WORK

The current scenario there is no UNL dictionary for Tamil language in full form. We attempted to create a vocabularies related to academia. The ontology for the same will also be created for correct translation in to Tamil. For the time being we have concentrated only Tamil Nadu Universities web sites. The domain specific words of academic sites have been populated. We have achieved 150 words and they have to be arranged in the hierarchy of UNL ontology.

V. REALATED WORK

Emhimed Salem Alatrish et. al. have proposed a semi-automatic procedure to create ontologies for different natural languages[8]. It aims to integrate different software tools which provides the building of ontologies for different natural languages. Nitsan chrizman et. al. have focused o automatic construction of multi-lingual domain-ontologies[9]. In this work it aims to create DAG(Directed Acyclic Graph) which consists of the concepts related to a specific domain and the relations between them.

Hadhemi Achour et. al. have proposed semantic ontology based model for multilingual indexing and retrieving the educational resources of a web learning environment[10]. It uses a prototype to perform trilingual searching (Arabic-English-French) for online learning resources. It uses indexing a database of learning resources related to the theme of object oriented programming in Java of the domain of computer science.

REFERENCES

[1] Levesque, Hector; Ronald Brachman (1985). "A Fundamental Tradeoff in Knowledge Representation and Reasoning". In Ronald Brachman and Hector J. Levesque. Reading in Knowledge Representation. 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, Web based UNL Ontology Visualization, Journal of convergence information technology, volume 8, number 13, August 2013, pp:69-75.

[8] Emhimed Salem Alatrish, Dusan Tosic and Nikola Milenkovic,"Building ontologies for different natural languages", Computer Science and Information Systems 11(2):623-644, DOI:10.2298/CSISI30429023A, 2014

[9] Nitsan Chrizman and Alon Itai, "How to construct Multilingual Domain Ontologies",[online]. URL:www.irec-conf.org/proceedings/Irec2014/pdf.730_paper.pdf, 2014.

[10] Hendez, M., Achour, H. "Keywords extraction for automatic indexing of e-learning resources", computer Applications & Research (WSCAR), World Symposium on pp: 1-5, 2014.