Collaborative Work on Indonesian WordNet through Asian WordNet (AWN)

Hammam Riza
Agency for the Assessment and Application of Technology (BPPT), Indonesia
hammam@iptek.net.id

Budiono
Agency for the Assessment and Application of Technology (BPPT), Indonesia
budi@iptek.net.id

Chairil Hakim
Agency for the Assessment and Application of Technology (BPPT), Indonesia
chairil@iptek.net.id

Abstract

This paper describes collaborative work on developing Indonesian WordNet in the AsianWordNet (AWN). We will describe the method to develop for collaborative editing to review and complete the translation of synset. This paper aims to create linkage among Asian languages by adopting the concept of semantic relations and synset expressed in WordNet.

1 Introduction

Multilingual lexicons is of foremost importance for intercultural collaboration to take place, as multilingual lexicons are several multilingual application such as Machine Translation, terminology, multilingual computing.
WordNet is the resource used to identify shallow semantic features that can be attached to lexical units. The original WordNet is English WordNet proposed and developed at Princeton University WordNet (PWN) by using bilingual dictionary.
In the era of globalization, communication among languages becomes much more important. People has been hoping that natural language processing and speech processing. We can assist in smoothening the communication among people with different languages. However, especially for Indonesian language, there were only few researches in the past.
The Princeton WordNet is one of the semantically English lexical banks containing semantic relationships between words. Concept mapping is a process of organizing to forming meaningful relationships between them.
The goal of Indonesian AWN database management system is to share a multilingual lexical database of Indonesian language which are structured along the same lines as the AWN.
AWN is the result of the collaborative effort in creating an interconnected Wordnet for Asian languages. AWN provides a free and public platform for building and sharing among AWN. The distributed database system and user-friendly tools have been developed for user. AWN is easy to build and share.
This paper describes manual interpretation method of Indonesian for AWN. Based on web services architecture focusing on the particular cross-lingual distributed. We use collective intelligence approach to build this English equivalent. In this sequel, in section 2 the collaborations builders works on web interface at www.asianwordnet.org. In section 3, Interpretation of Indonesian AWN, short description of progress of English – Indonesian translation and the obstacle of translation.

2 Collaborative AWN

WordNet covers the vast majority of nouns, verbs, adjectives and adverbs from English language. The words are organized in synonym sets called synset. Each synset represents a concept includes an impressive number of semantic relations defined across concepts.
The information encoded in WordNet is used in several stages in the parsing process. For instance, attribute relations, adjective/adverb classifications, and others are semantic features extracted from WordNet and stored together with the words, so that they can be directly used by the semantic parser.

Proceedings of the 8th Workshop on Asian Language Resources, pages 9–13,
Beijing, China, 21-22 August 2010.

To build language WordNet there are two main of discussion; the merge approach and the expand approach. The merge approach is to build the taxonomies of the language (synset) using English equivalent words from bilingual dictionaries. The expand approach is to map translate local words the bilingual dictionaries. This approach show the relation between senses. The system manages the synset assignment according to the preferred score obtained from the revision process. For the result, the community will be accomplish into original form of WordNet database. The synset can generate a cross language result.

AWN also introduce a web-based collaborative workbench, for revising the result of synset assignment and provide a framework to create AWN via linkage through PWN synset. AWN enables to connect and collaborate among individual intelligence in order accomplish a text files.

At present, there are ten Asian language in the community. The amount of the translated synsets had been increased. Many language have collaboration in AWN.

- Agency for the Assessment and Application of Technology (BPPT), Indonesia
- National Institute of Information and Communications Technology (NICT), Japan
- Thai Computational Linguistics Laboratory (TCL), Thailand
- National Electronics and Computer Technology Center (NECTEC), Thailand
- National University of Mongolia (NUM), Mongolia
- Myanmar Computer Federation (MCF), Myanmar
- National Authority of Science and Technology (NAST), Lao PDR
- Madan Puraskar Pustakalaya (MPP), Nepal
- University of Colombo School of Computing (UCSC), SriLanka
- Vietnamese Academy of Science and Technology (VAST), Vietnam

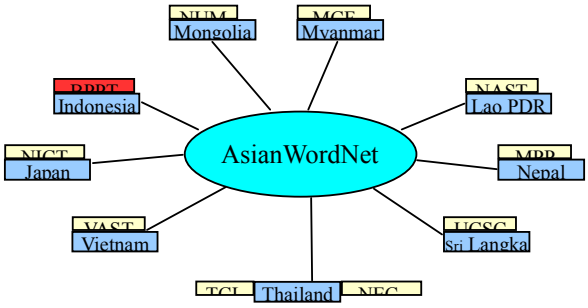

Fig 1. Collaboration on Asian WordNet

3 Interpretation of Indonesian AWN

Indonesian WordNet have been used as a general-purpose translation. Our approach was to generate the query for the web services engine in English and then to translate every key element of the query (topic, focus, keywords) into Indonesian without modifying the query. The dictionary is distinguished by set of entry word characteristic, clear definitions, its guidance on usage. All dictionary information for entries is structured such as entry word, multiple word entries, notes, contemporary definitions, derivations, example sentence, idioms, etc. All dictionary are implemented as text-files and as linguistic databases connected to Indonesian AWN. The set of language tags consists of part of speech, case, gender, number, tense, person, voice, aspect, mood, form, type, reflexive, animation.

3.1 Progress English – Indonesian

Indonesian WordNet is used Word Net Management System (WNMS) tools developed by AsianWordNet to create web services among Asia languages based on Princeton WordNet® version 3.0, Co-operation by TCL and BPPT establish on October 2007.

As presented above, we follow the merge to create and share the Indonesian WordNet by translating the each synonym translation. We expand an appropriate synset to a lexical entry by considering its English equivalent.

We plan to have reliable process to create and share Indonesian WordNet in AWN. We classify this work into four person AWN translators to participate in the project of Indonesian AWN.

Each person was given a target translator in a month should reach at least 3000 sense so that the total achievement 12000 senses in a month. From 117.659 senses that there is expected to be completed within 10 months. On the process of mapping, a unique word will be generated for every lexical entry which contain. The grammatical dictionaries contain normalized entry word with hyphenation paradigm plus grammatical tags.

	Assignment			TOTAL
	March	April	May	sense
Noun	10560	14199	16832	82115
verb	6444	6444	6499	13767
Adjective	1392	1392	1936	18156
Adverb	481	481	488	3621
Total	18877	22516	25755	117659

Table 1. Statistic of synsets

In the evaluation of our approach for synset assignment, we selected randomly sense from the the result of synset assignment to English – Indonesian dictionary for manually checking. The random set cover all types of part-of-speech. With the best information of English equivalents marked with CS=5. The word entry must be translated into the appropriate words by meaning explanation.
Table 1. presents total assignment translated words into Indonesian for the second third month. Following the manual to translate the English AWN to Indonesian, we resulted the progress of AWN at this time.
We start to translate or edit from some group of base type in "By Category". These base types are based on categories from PWN. There is only 21.89% (approved 25,755 of 117,659 senses) of the total number of the synsets that were able to be assigned to lexical entry in the Indonesian – English Dictionary.

3.2 Obstacle of Indonesian Translation

Wordnet has unique meaning of word which is presented in synonym set. Each synset has glossary which defines concept its representation. For examples word car, auto, automobile, and motorcar has one synset.

An automatic compilation of dictionary in AWN have a translational issues. There are many cases in explanation sense. One word in English will be translated into a lot of Indonesian words, glossary can be express more than one Indonesian word (Ex. 1).
One of the main obstacles in studying the absorption of English words in Indonesian words, is the fact that the original form of some words that have been removed due to Indonesian reform process, in which some words have been through an artificial process. There is no special character in Indonesian word, especially in technical word, so that means essentially the same as the English word (Ex. 2).

Ex. 1. **time frame**

POS	noun time
synset	time_frame
gloss	a time period during which something occurs or is expected to occur; an agreement can be reached in a reasonably short time frame"
Indonesian	jangka waktu, selang waktu

Ex. 2. **resolution**

POS	noun phenomenon
synset	resolution
gloss	(computer science) the number of pixels per square inch on a computer generated display; the greater the resolution, the better the picture
Indonesian	resolusi

Using definitions from the WordNet electronic lexical database. A major problem in natural language processing is that of lexical ambiguity, be it syntactic. Each single words must be container for some part of the linguistic knowledge need to ambiguous wordnet sense. Therefore,

not only a single heuristic translate Indonesian words. The WordNet defined in some semantic relations, this categories using lexicographer file and glossary definitions relations are assigned to weight in the range. WordNet hierarchy for the first sense of the word "*empty*" there are 10 synset words (I take three of ten) that are related to the meaning are the following in (Ex. 3.)
Three concepts recur in WordNet literature that entail a certain amount of ambiguity : terminological distance, semantic distance and conceptual distance. Terminological distance, by contrast, often appears to refer to the suitability of the word selected to express a given concept. Semantic distance is understood to mean the contextual factor of precision in meaning. And the conceptual distance between words, in which have relations proved.

Ex. 3. **empty**

POS	noun art
synset	empty
gloss	a container that has been emptied; "return all empties to the store"
Indonesian	hampa

empty

POS	verb change
synset	empty, discharge
gloss	become empty or void of its content; "The room emptied"
Indonesian	mengosongkan

empty

POS	adjectives all
synset	empty
gloss	emptied of emotion; "after the violent argument he felt empty"
Indonesian	kosong, penat

Disambiguation is unquestionably the most abundant and varied application. It precision and relevance in response to a query inconsistencies. Schematically the semantic disambiguation are selected in the glossaries of each noun, verb, and adjectives and its subordinates.
WordNet information, whose objective is to build designs for the association between sentences and coherence relations as well as to find lexical characteristics in coherence categories. WordNet became an ancillary tool for semantic ontology design geared to high quality information extraction from the web services.
A comparative analysis of trends in wordnet use :

1. Support for the design of grammatical categories designed to classify information by aspects and traits, but in particular to design and classify semantic ontologies.
2. Basis for the development of audio-visual and multi-media information retrieval systems.

4 Internet Viewer

The pilot internet service based on Wordnet 3.0 is published at http://id.asianwordnet.org.

5 Discussion and Conclusion

Any multilingual process such as cross-lingual information must involve resources and language pair. Language specific can be applied in parallel to achieve best result.
In this paper we describe manually sharing of Indonesian in the AWN by using dictionaries. AWN provides a free and public platform for building and sharing among AWN. We want continue the work defined learning the service matching system. Our future work on AWN will focuses in development platform WordNet and language technology web services.
Although AWN application are going steadily, the limitations are:

1. AWN designed for manual so authenticity can not be a reference.

2. Classification was performed manually, which means that the reasons and depth of classification may not be consistent.

References

Valenina Balkova, Andrey Suhonogov, Sergey Yablonsky. 2004. Rusian WordNet: From UML-notation to Internet/Infranet Database Implementation. In Porceedings of the Second International WordNet Conference (GWC 2004),

Riza, H., Budiono, Adiansya P., Henky M., (2008). I/ETS: Indonesian-English Machine Translation System using Collaborative P2P Corpus, Agency for the Assessment and Application of Technology (BPPT), Indonesia, University of North Texas.

Shi, Lei., Rada Mehalcea, (2005), Putting Pieces Together : Combining FrameNet, VerbNet, and WordNet for Robust Semantic Parsing

Thoongsup, S., Kergrit Robkop, Chumpol Mokarat, Tan Sinthurahat, (2009). Thai WordNet Construction. Thai Computational Linguistics Lab., Thailand

Virach Sornlertlamvanich., The 5th International Conference of the Global WordNet Association (GWC-2010), Mumbai, India , 31st Jan. - 4th Feb., 2010.

Fragos, Kostas, Yannis Maistros, Christos Skourlas, (2004). Word Sense Disambiguation using WORDNET relations. Dep. Of Computer Engineering NTUA, Greece.

www.asianwordnet.org