ELSEVIER

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 27 (2011) 188 – 193

Procedia
Social and Behavioral Sciences

Pacific Association for Computational Linguistics (PACLING 2011)

Statistical Malay Dependency Parser for Knowledge Acquisition Based on Word Dependency Relation

Hassan Mohamed[a], Nazlia Omar[a], Mohd Juzaidin Ab Aziz[a], Suhaimi Ab Rahman[b]

[a]*School of Computer Science, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
[b]*College of Information Technology, Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia*

Abstract

One of the common problems faced when processing information gathered from any natural language is the 'semantic gap' where the 'meaning' of the sentences is not exactly extracted. In Malay Natural Language Processing (NLP), as our knowledge, there is no existing Malay Parser that can be used to develop a knowledge acquisition feature to extract 'meaning' from Malay articles based-on syntactic relations. This relation is basically the relation between a word and its dependents. This paper will examine the Dependency Grammar (DG) for developing Malay Grammar Parser and discuss the possibilities of developing probabilistic dependency Malay parser using the projected syntactic relation from annotated English corpus. The English side of a parallel corpus, project the analysis to the second language (Malay). Thus, the rules for adaptation from English DG to Malay DG will be defined. The projected tree structure in Malay will be used in training a stochastic analyzer. The training will produce a set of tree lattices which contains chunks of dependency trees for Malay attached with their probability value. A decoder will be developed to test the lattices. A DG for a new Malay sentence is built by combining the pre-determined lattices according to their plausible highest probability of combination.

Keywords: Parser; Dependency Parser; Malay Parser; Syntactic Relation; Dependency Grammar; Malay corpus

1. Introduction

A system that extracts knowledge from texts indicates how the inferences necessary for the extraction of knowledge from sentences [1]. For example, one of the latest developments in engineering knowledge extraction [2] requires natural language processing tool to mark the relevant semantic annotations. The

doi:10.1016/j.sbspro.2011.10.597

natural language processing tool can bridge-up the gap between syntactic to semantic by the hope that the 'meaning' of sentences can be extracted. However, this is a problem when processing information gathered from any natural language. The 'meaning' of the sentences is not exactly extracted due to the 'semantic gap'. In order to extract knowledge from natural language sentences, one of the possible hidden information that needs to be examined is the syntactic relation. This relation can be seen as the relation between a word and its dependents. Therefore, before the problem for extracting knowledge from articles can be done, word dependencies need to be analyzed. For Malay language knowledge extraction, we need a Malay Dependency parser to analyze Malay sentence. This paper discusses the possibilities of developing Malay parser using the projected syntactic relation from annotated English corpus. The reason for looking into this approach is because English has its parser. Mainly, we will adapt the work done by Hwa et al. [4] by looking into parallel English – Malay corpus.

In order to develop high quality parsers, we need an annotated corpus with the desired linguistic representations basically known as "treebank" [4]. This effort is labor intensive and time-consuming process, and it is difficult to find linguistically annotated text in sufficient quantities. Therefore, there is a need to explore parallel text to help solving the problem of creating syntactic annotation in Malay language. The idea is to use the English side of a parallel corpus, project the analysis to the second language (Malay), and then train a stochastic analyzer. Therefore stochastic dependency parsers will be developed via projection from English. This paper discusses the possibilities of developing a Malay parser using the projected syntactic relation from annotated English corpus. The reason for looking into this approach is based on the fact that English has its parser and this will provide a sound ground for a head start in the under resourced Malay.

2. Related Work

Mosleh et al. [9] in developing the English – Malay Machine Translation (MT) used Synchronize Structured String Tree Correspondence (S-SSTC) to relate expressions of a natural language (source language) to its associated translation in another language (target language). S-SSTC is defined to make such relation explicit to facilitate such structural annotation to annotate the examples (translation units) in the Bilingual Knowledge Bank (BKB) [10]. The dependency structure has been chosen as the linguistic representation of the SSTC as it gives a natural way to establish the translation units between the *source* (English) and *target* (Malay) SSTCs. Fortunately, the S-SSTC representation schema was used for English – Turkish MT in Deniz [11] work.

3. Dependency Grammar

Dependency grammar (DG) is a class of syntactic theories developed by Lucien Tesnière [6]. The relation between a word which is head and its dependents determine the structure or dependency tree. Hence, DG is not determined by specific word order, but rather, concerned directly with individual words. Therefore, the DG is close to the 'meaning' because it tells about what companions a word can have by constructing an asymmetric head-modifier (governor-dependent) kind of relation.

Graphically, dependency trees can be represented with arrows pointing from the head to the dependents or from the dependents to the heads [7]. Fig 1 shows the graphical form of DG for an English sentence "The rabbit challenged the tortoise to race".

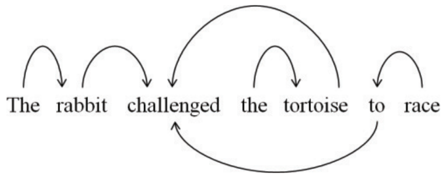

Fig 1. Dependency tree for the sentence “The rabbit challenged the tortoise to race”.

The arrows are moving from the dependents to their head, thus in Fig. 1, the word *rabbit*, *tortoise* and *to* are the dependents of *challenged*. This dependency can also be represented in textual form as shown in Table 1. The words of the sentence are in the second column, preceded by a column with word numbers. Further columns are added for the reference word number to indicate the dependencies. When the dependency reference number is zero means the word becomes the root or head.

Table 1. An example of a table

Word number	Words	Dependencies reference number
1	The	2
2	rabbit	3
3	challenged	0
4	the	5
5	tortoise	3
6	to	3
7	race	6

4. Projecting English to Malay

The analysis of Malay sentence is done through projecting from English sentence. For example consider a Malay sentence “Mariam memberikan Johan satu buku” which has an English translation “Mary gives John a book”. By using an English parser, the English sentence is analyzed as represented in graphical form for clearer picture to project to Malay side as in Fig. 2(a). The words in the dependency tree will be replaced by the translated Malay words with their own part of speech attached as shown in Fig. 3(b). Some semantic features are also added to the Malay part such as ‘subject:agent’, ‘Indirect Object:Beneficiary” and “Direct Object:Patient”.

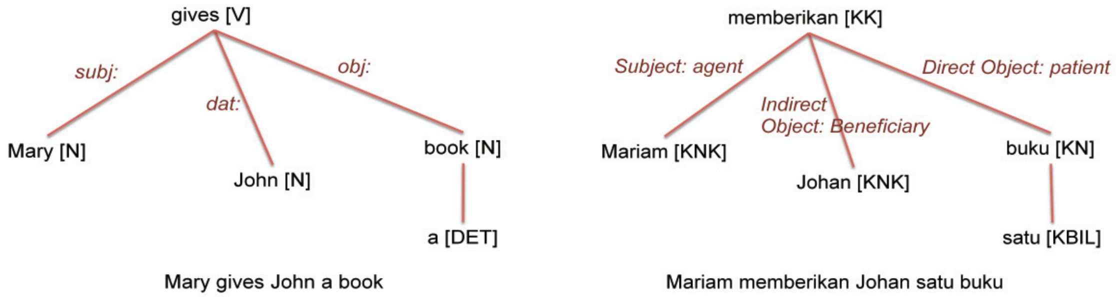

Fig. 2. (a) Dependency tree of “Mary gives John a book” sentence; (b) Dependency tree of “Mariam memberikan Johan satu buku” sentence

5. Proposed Research Design

The objective of this paper is to highlight an important issue on how to develop a Malay Dependency Grammar parser for knowledge acquisition or information extraction. We will adapt the work done by Hwa et al. [4] by looking into parallel English – Malay to make use of Direct Correspondence Assumption (DCA) and apply the pseudo DPA (PDPA) introduced by Goyal et al. [5] in projecting and filtering the source to the target. Furthermore we will adapt Spreyer et al. [3] work to investigate how graph-based and transition-based parser can benefit from the projection approach. As we adapt from [4] , the alignment is shown in Fig. 3(a). The overall architecture is shown in Fig. 3(b).

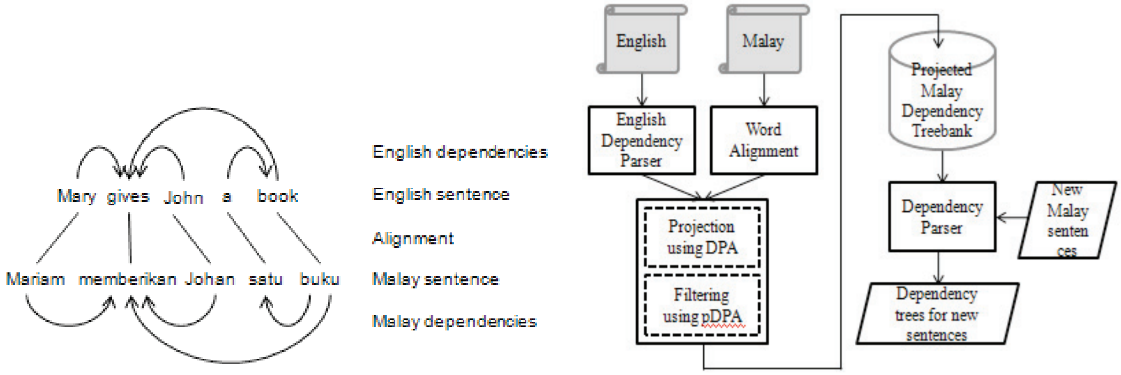

Fig. 3. (a) Projecting a dependency tree from English to Malay; (b) System architecture

5.1. Phase I: Data Collections and Preparation

This research needs collection of original Malay articles with English translation to be used for creating the parallel corpus in Phase II. The article will be divided based on their domains in order to have comparisons of the result versus their domains. Among the proposed articles that might be useful for this research are as follows: (1) General domain is taken from Malaysia Parliament Hansard and Kamus Inggeris Melayu Dewan (KIMD). The Hansard needs to be translated to English to be parallel but the KIMD is already in parallel version because there are a lot of sentence examples that explains the meaning of certain words. (2) Financial domain is taken from Bank Negara Malaysia reports which already exists in parallel version from yearly reports, quarterly reports, monthly reports and insurance/takafful reports. For a start, the target number of sentences for this research is about 30,000.

5.2. Phase II: Corpus Preparation

This research needs English – Malay parallel corpus for training and testing the stochastic model. The data must be annotated accordingly before they are used. In order to complete this phase, the following step will be taken: (a) English sentence will be parsed using available English Dependency Grammar parser; (b) once the dependency tree in (a) is produced, the structure will be projected to the translated language (Malay language); and then (c) the dependency tree for Malay will be edited to suit to the Malay syntactic rules.

In accordance to this phase, the rules for building Malay functional dependency will be defined with consultation from linguist experts. Some of these rules have been defined by a group of linguist during UKM-MIMOS 2006 project[1], but they still need to be added or revised to support this research purpose. Once the rules have been defined, the annotation process can proceed by editing Malay dependency structures through an editor with linguist's assistance and confirmation.

5.3. Phase III: Model Training

Stochastic analyzer is developed to train the model. Corpus produced in Phase II will be divided into 90% for training and 10% for testing. This training will produce tree lattices, which is chunks of Malay dependency tree with their probability value. These probability values are assigned by the stochastic analyzer after completing the training. The following formula will be used to count the probability [8].

$$P(\alpha \to \beta) = \frac{count\ (\alpha \to \beta)}{\sum_{\gamma} count\ (\alpha \to \gamma)}$$

$$= \frac{count\ (\alpha \to \beta)}{count\ (\alpha)}$$

5.4. Phase IV: Model Prototyping and Testing

A decoder is developed. The decoder will combine lattices to build a dependency tree for new sentences. The combination will be based on criteria where the most plausible combination is depending on the maximum probability value of certain dependency structure. The formula is as follows [8]:

$$\dot{T}(S) = \underset{T\, \in \tau(S)}{\operatorname{argmax}} P(T)$$

$$P(T) = \prod_{n\, \in T} P\big(l(n)\big)$$

T is the parse-tree of a sentence S. Lattices l are used to expand each node n in the parse-tree T. If the sentence has ambiguous T, $\dot{T}(S)$ will disambiguate them. The accuracy of parsing the Malay sentence will be evaluated based on the comparison between hand-crafted Malay DG and the DG produced by decoder. The formula used is as follows:

$$recall\ = \frac{no.of\ correct\ DG\ in\ candidate\ parse\ of\ S}{no.of\ correct\ DG\ in\ handcrafted\ parse\ of\ S}$$

$$precision = \frac{no.of\ correct\ DG\ in\ candidate\ parse\ of\ S}{no.of\ total\ DG\ in\ candidate\ parse\ of\ S}$$

6. Conclusion

As a respond to the challenge in information processing based on Malay text articles, the paper has

[1] Draft 4: Guidelines to Functional Dependency Grammar (FDG) for English and Malay Structures, UKM-MIMOS Team June 27, 2006. (Unpublished)

discussed the potential approach in developing Malay Dependency Parser. The parser is the basic tool for Natural Language Processing to be used in many field of NLP based applications such as information extraction, information retrieval, machine translation etc. Furthermore, to equip those applications with meaning-based features Dependency Grammar has been chosen for developing Malay parser and to leverage the existing English parser, the projected syntactic relation from annotated English has been used for Malay. Hence, the rules for adaptation from English DG to Malay DG can be developed. The projected tree structure in Malay will be used in training a stochastic analyzer. The training will produce a set of tree lattices which contains chunks of dependency trees for Malay attached with their probability value. A decoder will be developed to test the lattices. A DG for a new Malay sentence is built by combining the pre-determined lattices according to their plausible highest probability of combination.

References

[1] F. Gomez, “Semantic Interpretation and knowledge extraction”, Journal Knowledge-Based Systems Volume 20 Issue 1, Elsevier, Amsterdam Netherlands, Feb 2007.

[2] C. Ho, “Development of an engineering knowledge extraction framework”, ICCCI 2010 Part 1LNAI 6421, Springer-Verlag, Heidelberg Berlin, 2010, pp. 413-420.

[3] K. Spreyer , and J. Kuhn, “Data-driven dependency parsing of new languages using incomplete and noisy training data”, Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Boulder Colorado, June 2009, pp. 12-20.

[4] R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. “Bootstrapping parsers via syntactic projection across parallel texts”, Natural Language Engineering volume 11, 2005, pp. 311-325. 2005

[5] S. Goyal , N. Chatterjee, “Parsing aligned parallel corpus by projecting syntactic relations from annotated source corpus”, Proceedings of the COLING/ACL on Main conference poster sessions, Sydney Australia, July 17-18, 2006, pp. 301-308.

[6] http://en.wikipedia.org/wiki/Dependency_grammar

[7] http://www.ilc.cnr.it/EAGLES96/segsasg1/node44.html

[8] D. Jurafsky and J. H. Martin, Speech and Language Processing, Prentice Hall, New Jersey USA, 2009.

[9] M. H. Al-Adhaileh, and T. E. Kong and Z. Yusoff, “A Synchronization Structure Of SSTC And Its Applications In Machine Translation”, Proceedings of the 2002 COLING workshop on Machine translation in Asia - Volume 16.

[10] M. H. Al-Adhaileh, T.H. Kong, “Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema”, MT Summit VII 1999, pp. 244-249.

[11] N. Deniz ALP, Ç. Turhan, “English to Turkish Example-based Machine Translation with Synchronous SSTC”, Fifth International Conference on Information Technology: New Generations, IEEE, 2008, pp. 674-679.