A Cross-lingual Part-of-Speech Tagging for Malay Language

Norshuhani Zamin[1] and Zainab Abu Bakar[2]
[1]*Faculty of Science and Information Technology,Universiti Teknologi PETRONAS, Perak, Malaysia*
[2]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor, Malaysia*

Keywords: Annotation Projection, Part-of-Speech Tagging, Malay Corpus, Dice Coefficient, Bigram Similarity.

Abstract: Cross-lingual annotation projection methods can benefit from rich-resourced languages to improve the performance of Natural Language Processing (NLP) tasks in less-resourced languages. In this research, Malay is experimented as the less-resourced language and English is experimented as the rich-resourced language. The research is proposed to reduce the deadlock in Malay computational linguistic research due to the shortage of Malay tools and annotated corpus by exploiting state-of-the-art English tools. This paper proposed a cross-lingual annotation projection based on word alignment of two languages with syntactical differences. A word alignment method known as MEWA (Malay-English Word Aligner) that integrates a Dice Coefficient and bigram string similarity measure is proposed. MEWA is experimented to automatically induced annotations using a Malay test collection on terrorism and an identified English tool. In the POS annotation projection experiment, the algorithm achieved accuracy rate of 79%.

1 INTRODUCTION

POS tagging is the first stage in automated text mining. The development of language technologies can scarcely begin without this initial phase. Automated POS tagging aims to label or tag each word in the text with its correct POS tag to reflect its syntactic and morphological categories. The term "POS tag" is often used interchangeably with the term "category", "word class" or "lexical category" in many linguistic publications. The POS tags are usually represented by some abbreviations to indicate word categories. The research outcome is exemplified in Figure 1.

Input : *Saya datang dari Malaysia*.
(Translation: I come from Malaysia.)
Output : *Saya* /PRP *datang* /VBD *dari*/ IN *Malaysia*/NNP
(NOTE: PRP = Pronoun, VBD = Verb, IN = Preposition, NNP = Proper Noun)

Figure 1: Example Outcome of Proposed Research.

Text-processing tools such as lemmatisers, POS-taggers, analyzers, stemmers and parsers are available for only a few rich-resourced languages, such as English, German and Japanese. As for the existing English POS-taggers, there are many advanced versions that have reached almost up to 98% accuracy with almost possibility for further enhancement (Tsuruoka et al., 2005; Indurkhya and Damerau, 2010).

Malay is a language spoken by approximately 300,000,000 users in Malaysia, Brunei, Singapore and Indonesia (El-Imam and Don, 2005). It is a type of Indo-European language. A corpus is a collection of various written or spoken texts in machine-readable forms and is usually unstructured. A corpus plays a vital role in NLP text-based research. Malay has relatively few resources and limited collection of corpus. This limitation has been a hurdle for Malay linguists to investigate the language computationally (Hassan, 1974).

Malay is inflectional language in which the language performs massive affixation, reduplication and composition (Tan, 2003). The uniqueness of these characteristics attracts linguists to explore the underlying challenges and opportunities of the Malay language. There are three types of word derivation in Malay – noun derivation, verb derivation and adjective derivation. New words are created by attaching affixes onto a root word. Some of the derived words remain the same meaning with the root words and some are totally different in meaning. There are four types of affixes in Malay – prefixes, suffixes, circumfixes and infixes (Sharum et al., 2010). This inflectional property gives a unique

Zamin N. and Abu Bakar Z..
A Cross-lingual Part-of-Speech Tagging for Malay Language.
DOI: 10.5220/0005150602320240
In *Proceedings of the International Conference on Agents and Artificial Intelligence* (ICAART-2015), pages 232-240
ISBN: 978-989-758-074-1

identity to Malay morphology. Examples of noun, verb and adjective derivation for common affixes in Malay are listed in Table 1.

Table 1: Derivation of a New Malay Words using Affixes.

Malay Affix	Malay Word	English Word
ajar + ***an***	*ajaran*	*teaching*
bel + *ajar*	*belajar*	*to learn*
meng + *ajar*	*mengajar*	*to teach*
di + *ajar* (transitive)	*diajar*	*being taught*
di + *ajar* + ***kan*** (intransitive)	*diajarkan*	*being taught*
mem + ***pel*** + *ajar* + ***i***	*mempelajari*	*to study*
di + ***pel*** + *ajar* + ***i***	*dipelajari*	*being studied*
pel + *ajar*	*pelajar*	*student*
peng + *ajar*	*pengajar*	*teacher*
pel + *ajar* + ***an***	*pelajaran*	*subject / education*
peng + *ajar* + ***an***	*pengajaran*	*lesson / moral of the story*
pem + ***bel*** + *ajar* + ***an***	*pembelajaran*	*learning*
ter + *ajar*	*terajar*	*accidentally taught*
ter + ***pel*** + *ajar*	*terpelajar*	*well-educated*
ber + ***pel*** + *ajar* + ***an***	*berpelajaran*	*educated*

Unfortunately, annotated textual data in Malay are currently scarce. In Malaysia, there are few Malay corpora that have been made available. They are mainly developed for academic use and not publicly accessible. They consist of various genres from story books, novels, and from management studies to political issues. Examples of private data include the Malay Practical Grammar Corpus (Abdullah et al., 2004), the Dewan Bahasa Pustaka (DBP) Database Corpus[1], the Malay Corpus by Unit Terjemahan Melalui Komputer from the University Science of Malaysia (Ranaivo, 2004; Chuah and Yusoff, 2002) and, more recently, the Malay LEXicon (MALEX) (Don, 2010). The freely available Malay Concordance Project Corpus[2] is a collection of 3,000,000 words extracted from classical Malay texts, ones that are not related to this research domain. Limited access to such important sources of linguistic knowledge is a major hurdle in Malay NLP research.

Annotating the corpus manually is a laborious and expensive task. This task is called corpus annotation in linguistic. Some research do provides annotation standards and guidelines for natural language annotation in various domains such as the research by Galescu and Blaylock (2012) and Roberts et al. (2009). Annotations can be of different nature, such as prosodic, semantic or historical annotation. The most common form of annotated corpora is the grammatically tagged one such as the POS tagged corpus. Annotation projection is a task within parallel corpora where information from Source Language (SL) is projected or map onto the Target Language (TL). Parallel corpora or bitext are extensively studied in the machine translation field where the aligned phrases and words are used to create translation models. A survey of the literature revealed that reuse of resources helps to reduce costs and overheads in system development (Mayobre, 1991; Bollinger and Pfleeger, 1990; Barnes and Bollinger, 1991; Kim and Stohr, 1998). This presents significant opportunities to leverage these pre-existing resources from a rich-resourced language, such as English, to avoid building new text-processing tools for a totally new language from scratch. The idea to exploit the linguistic information from one language to another is proposed.

Annotation projection using pre-aligned parallel corpus is demonstrated successfully in projecting coreference resolution in English-Portuguese parallel corpus (de Souza and Orăsan, 2011), relation detection in English-Korean parallel corpus (Kim et al., 2010), dependency analysis in English-Swahili parallel corpus (De Pauw et al., 2009), semantic roles in English-German parallel corpus (Padó and Lapata, 2005) and syntactic relations in English-Romanian parallel corpus (Mititelu and Ion, 2005).

2 PART-OF-SPEECH (POS) TAGGING

POS tagging is considered as an already solved problem in NLP study (Søgaard, 2010). The state-of-the-art in POS tagging accuracy is about 96%-97%, but for most Indo-European languages such as English, French, Dutch and German and not for others (Indurkhya and Damerau, 2010). There are three ways to POS tagging- unsupervised, semi-supervised and supervised. Christodoulopoulos et al., (2010) evaluates seven unsupervised POS taggers spanning nearly 20 years of work using Wall Street Journal (WSJ) corpus. All the existing seven unsupervised POS taggers were built using different techniques and algorithms.

The comprehensive review finds that many older algorithms perform well than the newer algorithms because the methods used in newer systems do not suit with POS induction task. Nevertheless, success-

[1] *http://www.dbp.gov.my*
[2] *http://mcp.anu.edu.au/*

ful rate in POS tagging is mostly contributed by the supervised and semi-supervised learning algorithms. For example the fully-supervised Stanford POS tagger (Toutanova et al., 2003), trained on the Wall Street Journal (WSJ) corpus, records 97.24% accuracy (Dickinson, 2007) and the work by Søgaard (2010) is an attempt at a semi-supervised POS tagger using Support Vector Machine (SVM), trained on POS-Tagged WSJ has successfully achieved 97% accuracy.

An error-driven transformation-based tagger for English, known as Brill's tagger, automatically learns and induces tagging rules from a pre-tagged English corpus (Brill, 1995; Brill, 1992). It is the first widely used tagger to have an accuracy of above 95%. A mirror of Brill's tagger is the latest version and is known as the CST tagger[3]. It is currently made available online.

The advancement of supervised learning algorithms so as to reach the level of inter-annotator agreement is due to the massive volume of labelled training data which serve as the input to the tagger. However, manual labelling of data is highly expensive and involves laborious preparatory work. On the other hand, an unsupervised POS tagger appears to be a more natural solution. Such a tagger does not make use of any pre-tagged corpora for training but only makes use of a sophisticated computational method to automatically induce word groups. One of the earliest POS tagging researches in this manner was from Merialdo (1994) who used Maximum Likelihood Model to train a trigram Hidden Markov Model (HMM) tagger. Banko and Moore (2004) proposed an improved framework using a discriminative model rather than an HMM. Goldwater and Griffith (2007) presented a Bayesian-based tagger and a standard trigram HMM with accuracy closer to most discriminative model.

In recent years, there have been a few attempts on the development of unsupervised POS taggers for less-studied languages, with accuracies reaching up to 76.1% (Christodoulopoulos et al., 2010). Hence, this study is conducted to help bridge the gap between the supervised and unsupervised algorithms.

3 EXISTING ENGLISH POS TAGGERS

The aim of this research is to overcome the shortage of NLP resources in Malay language by utilizing existing computational linguistic resources in other languages. As NLP has made rapid progress over the last decades, it seems realistic to apply an existing POS tagger to a bilingual corpus (English and Malay) as a 'bridge'. The resulting annotations are projected to the second, the resource poor language i.e. Malay using the proposed word alignment algorithm. The projection algorithm for resource induction requires an aligned parallel corpus or bilingual corpus.

As the initial phase, it is significant to test the performance of some existing POS taggers over the English terrorism corpus. The corpus has been previously created through the translation of Google Translate. It is the best statistical translator found thus far for Malay language. However, as the Google Translate often produces translation that contains grammatical errors, the assistance of human expert is required to correct the errors.

The comparative study involved three most advanced open-source POS taggers (i.e. CST, CLAWS and Stanford) over their capacity to recognize some standard word classes in the 25 selected news articles of Indonesia terrorism. These POS taggers are selected for the comparison because they are publicly available. Out of these three, the best POS tagger is chosen based on how accurate it performed compared to human annotations.

CST Tagger was built based on the methodology proposed by Brill (1992, 1995). It is said as the mirror image of the original Brill's Transformation-based Learning tagger. Brill's tagger was trained on Brown Corpus. The corpus was compiled in the 1960s at Brown University, Providence, Rhode Island as a general corpus (text collection) in the field of corpus linguistic (Kučera and Francis, 1967). Next candidate tool is the CLAWS tagger for English part-of-speech tagging[4]. This linguistic tool is developed by a group of linguists at the University of Lancaster, UK known as the University Centre for Computer Corpus Research on Language (UCREL). CLAWS stands for the Constituent Likelihood Automatic Word- Tagging System. CLAWS is the major UCREL's achievement. It is a stochastic tagger that uses 134 tagsets trained on human tagged British National Corpus (BNC)[5] of multiple genres with accuracy rate between 96% and 97% (Garside, 1987; Garside and Smith, 1997; Leech et al., 1994). The system used the frequency statistics drawn from the BNC corpus.

The third tool in the comparative study is the Stanford POS Tagger[6]. It is a state-of-the-art sto-

[3] *http://cst.dk/online/pos_tagger/uk/*

[4] *http://ucrel.lancs.ac.uk/claws/*

[5] *http://www.natcorp.ox.ac.uk*

[6] *http://nlp.stanford.edu/software/tagger.shtml*

chastic tagger using the Maximum Entropy algorithm and trained on Wall Street Journal (WSJ) corpus, a collection multiple genres news articles with Penn Treebank[7] tagsets (Toutanova et al., 2003). Tagged sentences are extracted and split into training, development and test data sets. Stanford POS tagger recorded 97.24% of its best accuracy.

This comparative experiment processed 25 news articles which consist of 263 sentences and 5413 words and the results are presented in Table 2 followed by discussions and justifications. The evaluation metric used are Precision (P), Recall (R) and F1-Score (F1).

Table 2: A comparative study between CST, CLAWS and Stanford POS Tagger on English Terrorism Corpus.

	# Correct	# Missed	# Wrong	P	R	F1
CST	2353	275	317	0.88	0.80	0.84
CLAWS	1981	193	419	0.83	0.80	0.79
Stanford	2426	550	722	0.77	0.66	0.71

Although, most recent research in automated POS tagging have explored stochastic algorithm, the results show that the rule-based CST Tagger outperformed other stochastic taggers with 84% accuracy. Accuracy represents the percentage of correctly identified matches the proposed algorithm detects from the possible matches. The success of CST Tagger lies within the rule-based algorithm to tag unknown words. CST Tagger obtained high accuracy because the linguistic information is captured indirectly through its Transformation-based Error-driven Learning into a small number of simple non-stochastic rules. In this experiment, CST Tagger performed reasonably well on terrorism news articles probably because it was trained on the Brown Corpus which contains 500 samples of English-language texts, totalling roughly one million words, compiled from works published in the United States in 1961. Among these samples are sample texts on Political Science and Government Documents which might describe the cases, stories or issues on terrorisms.

4 MALAY-ENGLISH WORD ALIGNER (MEWA)

MEWA is a hybrid method using heuristic and probabilistic approach that aligns English to its corresponding Malay word, thus projects the part of-speech (POS) of the English word which was assigned by the off-the-shelf POS tagger to the newly aligned Malay word.

In detailed, MEWA uses a probabilistic approach i.e. N-gram scoring method for two characters which is commonly referred as bigram and is integrated with a heuristic approach, Dice Coefficient function (Dice, 1945) in order to calculate the probability distribution of letter sequences between Malay and English texts. This is a hybrid method which never been applied in any research involving the Malay corpus.

4.1 Dice Coefficient Function

Dice Coefficient is a heuristic alignment method which differs from statistical alignment. It uses specific associative measures rather than pure statistical measures. The function applies basic heuristic rule – a word pair are chosen as the aligned words if the pair has the highest co-occurrence score. It is a simple and intuitive estimation approach for word alignment (Moore, 2004). The Dice Coefficient Function is shown in Equation 1.

$$S_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2} \quad (1)$$

The Dice Coefficient is calculated by counting the number sentences where the word co-occur ($P_i Q_i$), the number of occurrences of the English word (P_i) and the number of occurrences of the Malay word (Q_i). The measure is always between 0 and 1. The use of the Dice Coefficient function in bitext alignment research is inspired by the research of Dien (2005).

4.2 Bigram String Similarity Measure

Bigram is a simple probabilistic method to measure the string closeness of two different texts and often generates good results. Bigram method performs well on languages of different structures and are widely implemented in much text-mining research as shown in Jiang and Liu (2010), Tsuruoka et al., (2005), Tan et al., (2002) including the pioneer of text projection research, Yarowsky et al., (2001). A bigram is also referred as the first-order Markov model as it looks one token into the past and a trigram is a second order Markov model while N-gram is the Nth order Markov model (Jurafsky and Martin, 2000). N-gram has been widely demonstrated in early research for word prediction (Jurafsky et al., 1994; Jurafsky et al., 1997). The successful of an N-gram model is measured by its perplexity. A better model has the lowest perplexity value.

[7] *http://www.cis.upenn.edu/~treebank*

In most of the word prediction research, by far, both the bigram and trigram approach show a great success. However, a bigram model is suggested if the training data is insufficient (Gibbon et al., 1997). A bigram based model is faster and smaller at matching two languages with dissimilar patterns (Dunning, 1994) and performs better than the sequence of morphemes algorithm as experimented in (Florian & Ngai, 2001). Based on these facts, the research chose to work on bigram probabilistic model without any prior evaluation done.

Additionally, bigram model of two characters is proposed to improve the limitations found in the bitext project research of Dien and Kiem (2003). Their research proposed a word alignment algorithm using Dice Coefficient function, a dictionary look-up and Vietnamese morphological analyzer to automatically align Vietnamese and English bilingual corpora. The position of the words in each pair of sentences is pre-aligned by an existing tool called GIZA++ (Och and Ney, 2000). The accuracy of word alignment of Vietnamese-English was approximately 87% (Dien, 2002). The English corpus is tagged using the fnTBL-toolkit (Florian and Ngai, 2001). The similarity of morphemes between the source word (English) and the target word (Vietnamese) retrieved from a dictionary look-up is calculated. A morpheme is the smallest meaningful unit of a language that cannot be further divided. For example, the word *unbreakable* consists of three morphemes: *un* (signifying *not*), *break* (the base word) and *able* (signifying *can be done*). In order to obtain the morpheme, morphological analyzer is required. They have generalized the Vietnamese POS tagsets against the tagsets used by the fnTBL-toolkit, the adopted English tagger.

To date, there is no relevant research found that is related to bitext alignment involving Malay except a research by Al-Adhaileh et al., (2009) and Nasharuddin et al., (2013). Al-Adhaileh et al., (2009) proposed a text alignment algorithm to align Malay and English text using Smooth Injective Map Recognizer (SIMR) and Geometric Segment Alignment (GSA) algorithm. Both the algorithms were successfully used to align French-English, Korean-English and Chinese-English in previous research. The hybrid methods were experimented on 100,000 words Malay-English extracted from books of multiple genres.

5 HOW DOES MEWA WORKS?

MEWA works by aligning one Malay word to the most probable translated English word at a time. The overall process flow of MEWA is depicted in Figure 2.

Consider an example of POS annotation projection in Malay. We start with a source text, a Malay sentence *Polis Indonesia mendakwa empat lelaki bersenjata.* The text is translated to English using Google Translate and then tagged using CST Tagger, we get *Indonesian/NNP police/NN accused/VBN four/CD armed/ VBN men/NNS*. MEWA uses both texts and information from Malay-English lexicon and produces a tagged Malay sentence. MEWA works by aligning one Malay word at a time. Let's illustrate the process to align the word *mendakwa*. D_{ME} is a Malay-English lexicon which consists of terrorism related words in Malay and its possible translation in English. Lexemes for *dakwa* are stored and their translation is the thesaurus of the word as exemplified in Figure 3.3. The main use of lexemes in the lexicon is to avoid the laborious process to lemmatize the massive inflectional words of the Malay language.

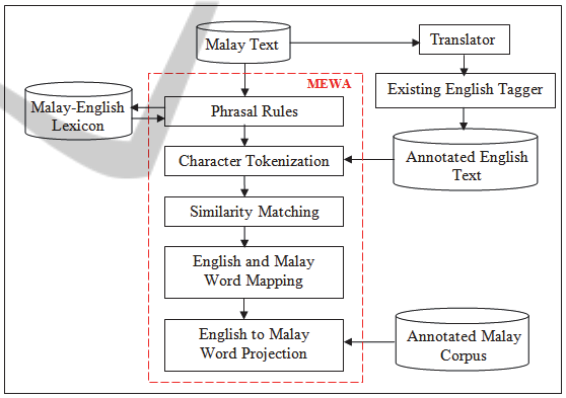

Figure 2: MEWA process flow.

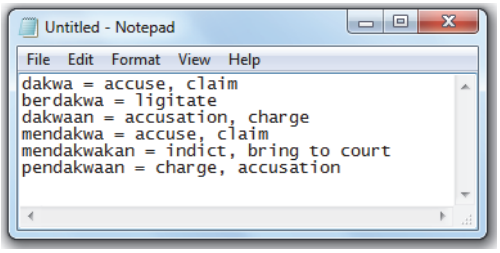

Figure 3: Sample lexicon entries for the Malay word *dakwa*.

Both the input texts are tokenized into Malay word vector, $W_{M,}$ and tagged English word vector, $W_{E,}$ as illustrated follows:

W_M = {*Polis, Indonesia, mendakwa, empat, lelaki, bersenjata*}
W_E= {*Indonesian/NNP, police/NN, accused/VBN, four/CD, armed/ VBN, men/NNS*}
= {E_1, E_2, E_3, E_4, E_5, E_6}

For each word in W_M we need to find the matching word in W_E. Taking a word from W_M, we use D_{ME} to find the possible English translations for the word – {d_i}. The morpheme similarity of each word d_i with each of the English words in W_E is now calculated with using the Sørensen-Dice function (Sørensen, 1948) as formulated in Equation 2.

$$Sim(M_i,E_j) = \max\{Sim\,(m_i\,, E_j)|m_i \in M_i, E_j \in W_E\} \quad (2)$$

Where M_i = A set of Malay vector; W_E = A set of tagged translated English word vector; m_i = The translation(s) of the Malay word; E_j = The translation(s) of the English word.

The lexicon search for the Malay word *mendakwa* finds the equivalent translation available from the manually prepared Malay-English lexicon. The word *mendakwa* is originated from the base word *dakwa*. From the lexicon, we get D_{ME}(*mendakwa*) = {*claim*, *accuse*} = {d_1, d_2}, where d_1 = *claim* and d_2 = *accuse*. Next, each of the words d_1 and d_2 is compared with each of the words E_1, E_2, E_3, E_4, E_5, and E_6. For example, the first iteration is to compare E_1 = *Indonesian* against d_1 = *claim* and d_2 = *accuse* and we get the following calculation:

Sim(*mendakwa*, *Indonesian*) = max{Sim(*claim*, *Indonesian*), Sim(*accuse*, *Indonesian*)}
= max{Sim({*cl,la,ai,im*}, {*In,nd,do,on,ne,es,si,ia,an*}), Sim({*ac,cc,cu,us,se*}, {*In,nd,do,on,ne,es,si,ia,an*})}
= max{Sim(2x0/(4+9)), Sim(2x0/(5+9))}
= max{0,0} = 0.

This calculation continues for the word E_2 (*police*) in the second iteration. The iteration continues until E_6 has been evaluated. The third iteration which compares E_3 = *accused* against d_1 = *claim* and d_2 = *accuse,* returns the highest correlation score as exemplified below:

Sim(*mendakwa,accused*)= max{Sim(*claim*, *accused*), Sim(*accuse*, *accused*)}
= max{Sim({*cl,la,ai,im*}, {*ac,cc,cu,us,se,ed*}), Sim({*ac,cc,cu,us,se*}, {*ac,cc,cu,us,se,ed*})}
= max{Sim(2x0/(4+6)), Sim(2x5/(5+6))}
= max{0,0.91} = 0.91.

In this example, the word *accused* is considered to be the most probable translation of the Malay word *mendakwa* as it returns the highest similarity score, 0.91. After all words in W_M and W_E are statistically compared, the bitext are mapped. Finally, the annotation i.e. the POS tag of each English word is projected to its corresponding Malay word so as to create the annotated Malay terrorism corpus, *Polis/NN Indonesia/NNP mendakwa/VBN empat/CD lelaki/NNS bersenjata/VBN*, as as shown in Figure 4.

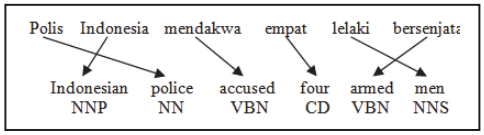

Figure 4: Malay-english bitext mapping.

6 RESULTS AND DISCUSSION

MEWA is evaluated using an English – Malay parallel corpus of 25 news articles on Indonesian terrorism, with 263 sentence pairs and 5413 word tokens, was used to evaluate the framework. The Malay texts were manually tagged by a human annotator using a standard tagsets, which were made equivalent to Brill's tagsets. The test tagsets are purposely reduced to six which refer to the broad categories of Brill's tagsets in order to simplify the human verification task. In this work, all variants of verbs, nouns, pronouns, cardinal numbers and adjectives produced by the CST Tagger are generalized as VB, CN, PN, CD and ADJ respectively to ease the evaluation process. These entities are significant for the later development. Regular words and delimiters were removed in the Malay corpus leaving only 3466 prominent word tokens. The experiment is divided into three sub-experiments:

- Experiment I: To test a small corpus of 50 randomly picked sentences from the collection of news articles.
- Experiment II: To test on a larger scale corpus consists of 263 sentences.

The result is shown in Table 3. In the Experiment II, a total of 263 sentences were extracted from a set of 25 Malay news articles on Indonesian terrorism. The system performance against human annotation results is shown in Table 4.

In both experiments, the numbers of correct, wrong and missed tags between the output produced by system and those annotated by our human experts were manually counted for all words. The results of the Experiment II demonstrated that the system achieved 86.87% in Precision, 72.56% in Recall and 79.07% in F1. These results indicate that the tagger attains slightly better annotation accuracy than in the Experiment I (refer to Table 3). The reason for this could be that the size of the corpus is 12 times larger than the previous size. The experiments show that the performance increases proportionally with the size of data. Observations have found that the size of

Table 3: Experiment I – Test result for 50 sentences.

# Sentences	# Word	# Correct	# Wrong	# Missed	P	R	F1
50	646	432	140	70	76 %	67%	71.2%

Table 4: Experiment II – Test result for 263 sentences.

# Sentences	# Word	# Correct	# Wrong	# Missed	P	R	F1
263	5413	2515	380	571	86.87%	72.56%	79.07%

the dictionary look-up also increases proportionally with the size of the tested corpus. Consequently, the probability of an English word getting correctly aligned with a Malay word is increased. A total of 1444 pairs of Malay-English words in the dictionary look-up have been collected from these experiments. The results indicate that this implementation works well when the sentence pair is good, i.e. when there is no data sparsity problem.

The performance was also contributed by the CST POS Tagger's expressive set of rules which is able to solve ambiguity problems in English. Ambiguity is among the challenging problems in word tagging (Dickinson, 2007). Tagging ambiguity is when there is a word with more than one POS tags exists. This occurs more frequently in English than Malay as in the example *We **can** **can** the **can***. The three occurrences of the word ***can*** correspond to the auxiliary, verb and noun categories respectively. As for Malay, there are limited words that carry different meanings in different contexts.

Consider this example: *Muzium **Perang** itu berwarna **perang***. In this example, the former ***perang*** means *war* while the latter means *brown*. The whole statement means *The War Museum is brown in colour*. This is an adjective clause where the latter ***perang*** is an adjective that describes the noun ***perang*** in the proper noun *Muzium Perang*. In the Experiment II, there were 571 missing words which fall under the "Missed" category. "Missed" means skipped or not processed by the system. It was found by observation that a word was missed due to either it not being in the dictionary / lexicon or it being associated with a many-to-one entity**.**

Clearly, system performance can be increased further by adding those missing words to the dictionary / lexicon. On the other hand, ambiguous tagging results returned by Brill's tagger contributed to reducing system performance. Figure 5 shows an erroneous tagging produced by the CST POS Tagger. Words given the wrong tag are underlined.

Dulmatin/NNP ,/, 39/NNP ,/, is/VBZ among/IN three/CD suspected/VBN militants/NNS who/WP were/VBD shot/NN dead/JJ

Figure 5: CST Tagger's tagging errors.

The word *suspected* in this context is supposed to be classified as adjective and given the tag *JJ* while the both the word *shot* and *dead* are tagged as *verb (VBD)* and *adjective (JJ)* respectively. However, ambiguous tagging results were found only in 112 words which is about 29% from the total wrong tagged words. Meanwhile, it is also observed in the dataset that there exist a conflict between the system's output and human's annotation in some word classes. As for example, the human annotators agreed that an adjective preceding a noun is also classified as a common noun (CN) as shown in Table 5.

Table 5: Contradiction between MEWA's tagging and Human's tagging.

Case	Examples
1	CST-tagged: anti-terror/JJ laws/NNS System-tagged: undang-undang/NNS antikeganasan/JJ Human-tagged: undang-undang/CN antikeganasan/CN
2	CST-tagged: mountainous/JJ areas/NNS System-tagged: kawasan/NNS pergunungan/JJ Human-tagged: kawasan/CN pergunungan/CN

7 CONCLUSIONS

A new method to automate POS-tagging for Malay using parallel data from a rich-resourced language is devised. As far as the research is concerned, no previous work has studied the alignment of a Malay corpus with an English corpus by projecting tags using hybrid algorithms. The bitext alignment method appears to be a powerful unsupervised learning algorithm mapping two dissimilar languages at minimum computational cost. MEWA has successfully mapped and project POS annotation from English to Malay corpus using existing English resource at fairly accurate rate. MEWA is the first attempt for cross lingual language research in Malay.

Finally, MEWA heavily reduces the labour required to annotate a Malay corpus, and generate quick results.

8 FUTURE WORKS

Limitation in this research is outlined in this section to be considered as topics to broach in our future research to improve the performance of MEWA. Some visible limitations that have been observed during the experiments are the disagreement with human's annotations. As shown in Table 4, the human annotators agreed that an adjective preceding a noun is also classified as a common noun (CN). However, the existing English tagger keeps it as adjective due the semantic of the word that uniquely described the noun. This contradict piece of categorization has decreased the system's performance.

Difficulties are identified in one-to-many, many-to-one and many-to-many word association. This is a common problems link with Machine Translation (MT) research for many years. A many-to-one association is the most common kind of association where an object can be associated with multiple objects. However, these problems can be resolved using rules. For a complete application, it is recommended to put the focus on the automated lexicon builder. The compilation of such lexicon (interchangeable referred to as dictionary look-up or gazetteer) is often a stumbling block in Natural Language Processing (NLP) research. Machine learning methods especially the rule-based algorithm is often used to perform this process. An advanced algorithm using the words extracted from Wikipedia has been greatly explored recently.

REFERENCES

Abdullah, I. H., Ahmad, Z., Ghani, R. A., Jalaludin, N. H., & Aman, I. (2004). A Practical Grammar of Malay–A Corpus based Algorithm to the Description of Malay: Extending the Possibilities for Endless and Lifelong Language Learning. *National University of Singapore*.

Banko, M., Brill, E. (2001). Mitigating the Paucity-of-Data Problem: Exploring the E□ect of Training Corpus Size on Classifier Performance for Natural Language Processing. *Proceedings of the first international conference on Human language technology research*, 1-5. Association for Computational Linguistics.

Barnes, B. H. & Bollinger, T. B. (1991). Making Reuse Cost-Effective. *IEEE Software,* 8(1), 13-24.Kim, Y., & Stohr, E. A. (1998). Software Reuse: Survey and Research Directions. *Journal of Management Information Systems*, 113-147.

Bollinger, T. B. & Pfleeger, S. L. (1990). Economics of Software Reuse: Issues and Alternatives. *Information and Software Technology*, 32 (10), 643-652.

Brill, E. (1992). A Simple Rule-based Part of Speech Tagger. In *Proceedings of the Workshop on Speech and Natural Language*, 112-116. Association for Computational Linguistics.

Brill, E. (1995). Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4), 543-565.

Christodoulopoulus, C., Goldwater, S., & Steedman, M. (2010). Two Decades of Unsupervised POS Induction: How Far Have We Come. *Proceedings of Empirical Methods in Natural Language Processing*.

Chuah C. K. & Yusoff, Z. (2002). Computational Linguistics at Universiti Sains Malaysia. *Proceedings of the International Conference on Language Resources and Evaluation***,** 1838 -1842.

De Pauw, G., Wagacha, P. W., & De Schryver, G. M. (2009). The SAWA Corpus: A Parallel Corpus English-Swahili. *Proceedings of the First Workshop on Language Technologies for African Languages*, 9-16. Association for Computational Linguistics.

De Souza, J. G. C., & Orăsan, C. (2011). Can Projected Chains in Parallel Corpora Help Coreference Resolution? *Anaphora Processing and Applications*, 59-69). Springer Berlin Heidelberg.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association between Species. *Ecology*, *26*(3), 297-302.

Dickinson, M. (2007). Determining Ambiguity Classes for Part-of-Speech Tagging. *Proceedings of RANLP-07. Borovets, Bulgaria*.

Dien, D. (2002). Building a Training Corpus for Word Sense Disambiguation in English-to-Vietnamese Machine Translation. In *Proceedings of the 2002 COLING Workshop on Machine Translation in Asia - Volume 16* (1-7). Association for Computational Linguistics.

Dien, D. I. N. H. (2005). Building an Annotated English-Vietnamese Parallel Corpus. *MKS: A Journal of Southeast Asian Linguistics and Languages*, 35, 21-36.

Dien, D., & Kiem, H. (2003). POS-tagger for English-Vietnamese Bilingual Corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using Parallel Texts: Data Driven Machine Translation and Beyond – Volume,* 88-95. Association for Computational Linguistics.

Don., Z. M. (2010). Processing Natural Malay Texts: a Data Driven Algorithm, *Journal of TRAMES*, 14, 90-103.

Dunning, T. (1994). *Statistical Identification of Language*, Computing Research Laboratory, New Mexico State University.

El-Imam, Y. A. & Don, Z. M. (2005). Rules and algorithms for phonetic transcription of standard Malay. *IEICE Transaction of Information Systems*, E88-D, 2354-2372.

Florian, R., & Ngai, G. (2001). Fast transformation-based Learning Toolkit, Technical Report.

Galescu, L., & Blaylock, N. (2012). A Corpus of Clinical Narratives Annotated with Temporal Information. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 715-720, ACM.

Galescu, L., & Blaylock, N. (2012). A Corpus of Clinical Narratives Annotated with Temporal Information. *Pro-*

ceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 715-720, ACM.

Garside, R. (1987). The CLAWS Word-Tagging System. *The Computational Analysis of English: A Corpus-based Algorithm. London: Longman*, 30-41.

Garside, R., & Smith, N. (1997). A Hybrid Grammatical Tagger: CLAWS4. *Corpus Annotation: Linguistic Information from Computer Text Corpora*, 102-121.

Gibbon, D., Moore, R. K., & Winski, R. (Eds.). (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Walter de Gruyter.

Goldwater, S., & Griffiths, T. (2007). A Fully Bayesian Algorithm to Unsupervised Part-of-Speech Tagging. In *Annual Meeting-Association for Computational Linguistics*, 45(1),744.

Hassan, A. (1974). The Morphology of Malay, Dewan Bahasa dan Pustaka, Kuala Lumpur Malaysia.

Indurkhya, N. & Damerau, F.J. (2010). *Handbook of Natural Language Processing, Second Edition,* Chapman & Hall / CRC Press.

Jiang, W., & Liu, Q. (2010). Dependency Parsing and Projection based on Word-pair Classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 12-20). Association for Computational Linguistics.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing*, Computational Linguistics and Speech, Prentice Hall.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., & Ess-Dykema, V. (1997). Automatic Detection of Discourse Structure for Speech Recognition and Understanding. *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on* (88-95). IEEE.

Jurafsky, D., Wooters, C., Tajchman, G., Segal, J., Stolcke, A., Foster, E., & Morgan, N. (1994). The Berkeley Restaurant Project. *ICSLP* (94,2139-2142).

Kim, S., Jeong, M., Lee, J., & Lee, G. G. (2010). A Cross-Lingual Annotation Projection Algorithm for Relation Detection. *Proceedings of the 23rd International Conference on Computational Linguistics*, 564-571. Association for Computational Linguistics.

Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Dartmouth Publishing Group.

Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: The Tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics,*1(622-628). Association for Computational Linguistics.

Mayobre, G. (1991). *Using Code Reusability Analysis to Identify Reusable Components from the Software Related to an Application Domain*. Proceedings of the 4th Annual Workshop on Software Reuse, 1-14.

Merialdo, B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, *20*(2), 155-171.

Mititelu, V. B., & Ion, R. (2005). Cross-Language Transfer of Syntactic Relations Using Parallel Corpora. *Cross-Language Knowledge Induction Workshop, Romania*.

Moore, R. C. (2004). Improving IBM Word-Alignment Model 1. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (518). Association for Computational Linguistics.

Och, F. J., & Ney, H. (2000). Giza++: Training of Statistical Translation Models.

Ranaivo, B. (2004). *Methodology for Compiling and Preparing Malay Corpus*. Technical Report. Unit Terjemahan Melalui Komputer. Pusat Pengajian Sains Komputer, Universiti Sains Malaysia.

Sharum, M. Y., Abdullah, M. T., Sulaiman, M. N., Murad, M. A., & Hamzah, Z. (2010). MALIM—A New Computational Algorithm of Malay Morphology. *Proceedings of Information Technology (ITSim),* 2, 837-843.

Søgaard, A. (2010, July). Simple Semi-Supervised Training of Part-of-Speech Taggers. *Proceedings of the ACL 2010 Conference Short Papers*, 205-208. Association for Computational Linguistics.

Sørensen, T. (1948). {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}. *Biol. skr.*, *5*, 1-34.

Tan, C. M., Wang, Y. F., & Lee, C. D. (2002). The Use of Bigrams to Enhance Text Categorization. *Information Processing & Management*, *38*(4), 529-546.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* 1,173-180. Association for Computational Linguistics.

Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. I. (2005). Developing A Robust Part-of-Speech Tagger for Biomedical Text. *Advances in Informatics*, 382-392. Springer Berlin Heidelberg

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. *Proceedings of the Human Language Technology Research,* 1-8.