NERSIL: the Named-Entity Recognition System for Iban Language

Soo-Fong Yong[a], Bali Ranaivo-Malançon[b], Alvin Yeo Wee[b]

[a]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak
94300 Kota Samarahan, Sarawak, Malaysia
soofong0629@gmail.com

[b]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak
94300 Kota Samarahan, Sarawak, Malaysia
{mbranaivo, alvin}@fit.unimas.my

Abstract. This paper presents NERSIL, the first Iban named entity recognition using ANNIE, an information extraction tool available within GATE. We proposed a method for building rules and gazetteers for Iban language. Rules were determined based on named entities that were not recognized by ANNIE. Then, the investigation of the contexts of these non-recognized Iban named entities allowed us to write the rules. NERSIL achieves 76.4% F-measure.

Keywords: Named Entity Recognition, GATE, Information Extraction, Rules-based, Gazetteers

1 Introduction

Named entities (NEs) are recognized as an important source of information for many applications. In information retrieval, they improve the detection of relevant documents. In machine translation, post-editing is very expensive when the errors of a machine translation system are mainly due to NEs. Named entity recognition (NER) is the process of detecting NEs and classifying them into semantic categories. NERSIL is the first Iban NER. One of the policies of the Malaysian government is to develop more Malaysian digital contents by including indigenous languages like Iban. Presently, NER can accommodate only major languages such as English and Chinese. Indigenous languages are largely neglected. Hence, NER which can accommodate indigenous languages such as Iban is required. Besides the development of the tool itself, our project will provide access to indigenous knowledge. These may in turn encourage more research, which in long term will preserve the culture. In addition, the proposed method may help other under-resourced languages.

The paper is structured as follows. Section 2 presents the background and related works. Section 3 explains the proposed method for building Iban rules and gazetteers. Section 4 describes the experiment performed to evaluate NERSIL, and section 5 depicts the analysis of the errors. Section 6 concludes the presentation and highlights our plans for future work.

2 Background and Related Works

2.1 Iban language

According to *Dewan Bahasa dan Pustaka* (Malay for The Institute of Language and Literature) there are 63 indigenous languages in Sarawak, an east Malaysian state. Iban is the largest ethnic group making up about 44% of the population of Sarawak. Iban language is the vernacular for Iban

people. The language is written using Latin alphabet. In 2008, Iban language was introduced as a new elective subject in the national fifth form examination.

The digitization and content processing of Iban texts are parts of our center research objective. Currently, we have already developed for Iban, a morphological analyzer and generator, a syntactic parser, a POS tagger, a spell checker, and an interactive alignment tool. The development of the Iban NER is in line with the research center objective. Even though a few processing tools have been developed for Iban, this language is still considered as an under-resourced language.

2.2 NER

NER process is divided into two successive parts. The first part consists of identifying proper names in a given text. The second part concerns the classification of these proper names into semantic categories such as Person, Organization, Location, Date, Time, Percentage, Monetary value and etc.

ANNIE is an information extraction tool widely used to identify NEs in different domains and languages (e.g., Maynard et al., 2003). ANNIE adopts a rule-based approach in recognizing NEs. A rule-based model consists of patterns (or rules) and gazetteers. Rules are manually written by language engineers. A gazetteer contains a list of words of the same kind of NE. Therefore, each type of NE will have its own gazetteer. In the literature, research on NER task for under-resourced languages is rare. In 2003, Maynard and her colleagues (Maynard et al., 2003) reported the development and evaluation of the first NER for Cebuano, a language used in the Philippines. They adapted ANNIE for Cebuano and achieved an F-measure of 77.5%. The authors reused the default setting (tokeniser, sentence splitter, POS tagger, gazetteer, NE grammar, and orthomatcher) and made some modifications by providing the Cebuano post-processor, Cebuano lexicon as well as Cebuano gazetteers. As ANNIE for Cebuano achieved high accuracy result, hence, it gives us some motivation to adapt also ANNIE for Iban language.

One of the difficulties of NER task is the identification of named entity variants. A named entity may be displayed in different formats. For instance, a date can be displayed as “10 Oct.”, “10 October 2010”, or “10/10/10”. Another difficulty of NER task is language ambiguity. A named entity may belong to more than one semantic category. In “Malaysia won the Suzuki Cup 2010”, “Malaysia” is considered as an organization. In “The Suzuki cup 2010 took place in Malaysia”, “Malaysia” is a location. In addition, style, structure, domain, genre, punctuation, capitalization, spelling, spacing as well as formatting also can be considered as NER difficulties.

3 Proposed Method

Given that there is no large corpus available for the indigenous languages we are working with, we attempted to develop the Iban NER tool as depicted in Figure 1.

To determine Iban rules for the automatic recognition of NEs, we first look for missing rules. We said that a rule is missing in ANNIE, if a NE that occurs in the Iban texts has not been identified by ANNIE. The main purpose of this experiment is to highlight the NEs that can be recognized by ANNIE. Next, the NEs are investigated manually. The result of this investigation is the creation of a set of Iban rules and gazetteers.

We tested the accuracy of NERSIL by comparing its results against human annotated texts, which have been labeled by local Iban native speaker. They used the “Simple Named Entity Guidelines V6.5” as a reference. This guideline has been designed specifically for less commonly

taught languages (LDC, 2006). Three metrics, available in GATE, were used for this evaluation: recall, precision, and F-measure.

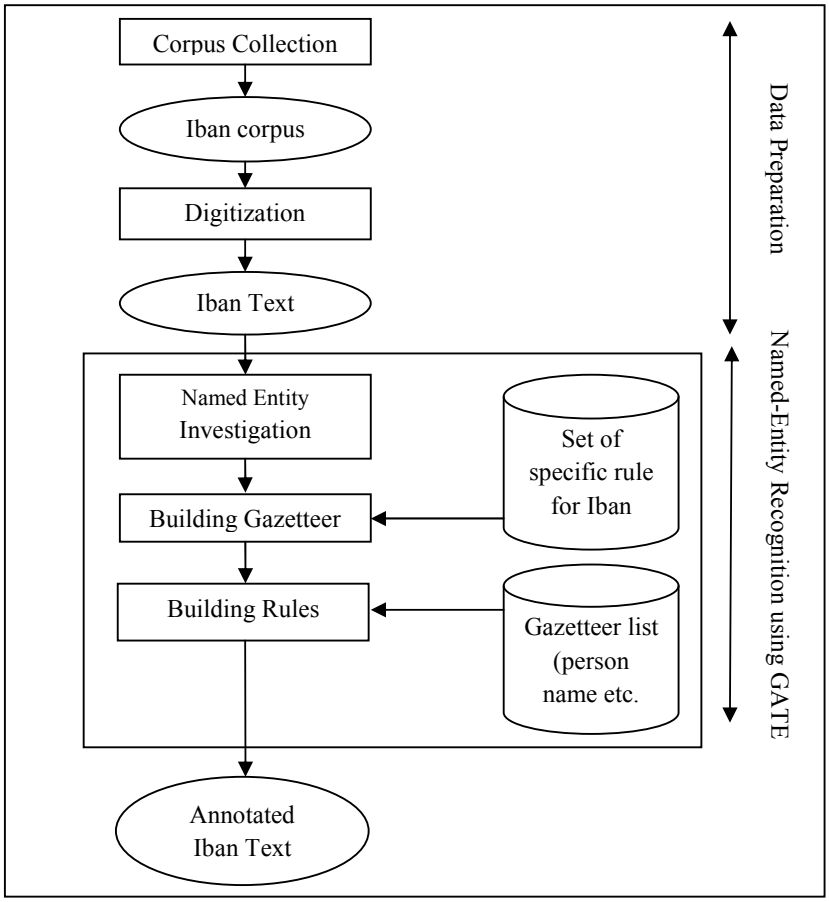

Figure 1: Proposed Method

Step1: Corpus Collection
The Iban texts corpus was collected from the Tun Jugah Foundation, which is one of the organizations carrying out research projects and activities on oral history and literature of Iban community.

Step2: Digitization
Since our Iban texts corpus is in paper format, we need to digitize it. We used as Optical Character Recognition (OCR) such as *ABBYY FineReader* software, which is able to recognize Malay texts. As Iban and Malay are close languages (Ng et al, 2010). We assume that this software will convert Iban corpus into electronic files with minimal errors.

Step3: Named Entity Investigation
In order to determine the best rules for the recognition of Iban NEs, we investigated the named entity of Iban. A known problem in the area of NER was found in our corpus that is cascaded NEs. This is the case when a named entity is embedded in another named entity. For instance, the whole

expression “Tuai CID Sibu [location] DSP Wong Chee Kiong” is tagged as a [Person]. However, “Sibu” will be tagged as a Location, “Wong Chee Kiong” as Person, and “Tuai CID” and “DSP” as Title. To make sure there are not duplicate rules happen, we can set the priority.

Step4: Building Gazetteers

The gazetteers are manually built. Although a somewhat tedious process, this was the only way to obtain accurate results. Besides that, due to the texts corpus is paper format. Thus, we have to manually check the validity. The gazetteers are the lookup list of NEs. During our corpus investigation, we have found that, like many language, Iban incorporates foreign words in its communication. These foreign words are mainly borrowed from English vocabulary. Therefore, we kept the English gazetteers available in ANNIE to ensure the recognition of all NEs, either native or foreign words, in Iban texts.

Step5: Building Rules

Rules in ANNIE are written using JAPE (Java Annotation Pattern Engine) (Maynard, 2011). Iban rules are built based on some contexts (Table2) and morphological features (Table3).

a. Person Rule

The identification of the entity Person, makes use of title information whether they are Malay titles (*Dato', Tan Sri*) or English Title (Dr., Mr.). Then, it checks the initial case status of the following words. A sequence of words is labeled as Person if it starts with a title followed by capitalized words. For example, “Tan Sri Dato” Sri Dr Lim Kok Wing” will be recognized as a Person. Knowing that “Tan Sri”, “Dato’ Sri”, and “Dr.” are titles, any occurrence of “Lim Kok Wing” will be also tagged as Person. Table 1 shows some examples of rules for Person identification.

Each culture has its own system of naming. For example, Malay names have the following pattern: (<name> bin or binti <father’s name>, where “bin” is for men or “binti” is for women, e.g., Musa bin Osman); Chinese names, as used in Malaysia, have three segments: (<surname>< name1><name2>, e.g., Chong Ai Ling); Indian names in Malaysia have similar pattern as Malay names. The difference is on the words used to express the relationship with the father. If Malay uses “bin” and “binti”, Indian uses “a/l”, for “anak lelaki” (“son of”) or a/p, for “anak perempuan” (“daughter of”), e.g., Anbuselvan a/l Ramanan Similarly, the names of Bumiputeras in Sarawak differ only on the middle word used in indicate the father’s name. They use anak (“child of”), as in Jugah anak Rentap. When one look closely to all these patterns, they are quite similar. The difference is on the words used to introduce the father’s name. Thus, we built a JAPE rule that can recognize the pattern <Name><Father’s affiliation><Father’s name>. By referring to the naming patterns, the naming patterns are quite similar. Thus, the naming patterns rules can simplified as generic rules, such as <Proper Noun><Unknown><Proper Noun>. We can try and predict the “unknown” words.

Table 1: Example of rules (Person)

Example of Rule	Explanation
Rule: Person	
({Token.kind !=”number”})	Check whether it is number, to avoid address pattern is tagged.
(	
({Token.kind==word, Token.orth==upperInitial})	At least one Capital letter word
({Token.kind==word, Token.orth==upperInitial})?	? refer to Capital letter word exist anot
({Token.kind==word, Token.orth==upperInitial})?	? refer to Capital letter word exist anot
):label	
({Token.string==”,”})	
({Token.kind==”number”})	Number refer to the age
({Token.string==”;”) +	+ mean that repeat the pattern again
→	Match the LHS rules with RHS
:label.Person={rule=”Person”}	label as Person

Table 2: List of contexts features for Iban (Budi, 2005)

Feature Name	Explanation	Example
PPRE	Person prefix	Tan, Teo (surnames)
PMID	Person middle	bin, b. binti, bt. (Malay names); a/l, a/p (Indian names); a/k, anak (Bumiputera names)
PTIT	Person title	Sapit Kepala Menteri Datuk Patinggi Tan Sri
OPRE	Organization prefix	Sekula/Sekula (School); Gerempong/ Gerempung (Community, association, cooperative); Kelab (Club); kunsil (council, Sarawak Administrative Officer); Kompeni (Company)
OSUF	Organization suffix	Sdn. Bhd /Sendirian Berhad (Incorporated); Beng/Bing (Bank)
OPOS	Position in organization	Tuai (Chief, head, leader)
LPRE	Location prefix	Jalai (road); Kampong (village); Bukit, Gunong (Mountain); Negeri/Nengeri (city); menoa/menua (settlement); Long (river mouth); Palan (resting place, camp, port)
POLP	Preposition that’s usually followed by person	Oleh (by) , untuk (for)
LOPP	Preposition that’s usually followed by location name	ba (at, in, on); di (at, in, on); dalam (in); ari (from); ngagai (towards); ke (for, to); enggau (with, together, and)
DAY	Day	Hari Senin (Monday)
MONTH	Month	Januari, Februari, Mac, Mei, Jun, Julai, Ogos, Oktober, Disember

Table 3: List of morphological features (Budi, 2005)

Feature Name	Explanation	Example
TitleCase	Begin with uppercase letter and followed by all lowercase letter	Soedirman
UpperCase	All uppercase letter	SHELL
LowerCase	All lowercase letter	Lemai (afternoon, evening)
MixedCase	Uppercase and lowercase letter	SaLT
CapStart	Begin with uppercase letter	-
ChartDigit	Letter and number	P3K
DigitSlash	Number with slash	12/5
Numeric	Number with dot or comma	20.5; 15,000
NumStr	Number in word	Sa/Satu (One)
Roman	Roman number	VII, XI
TimeForm	Number in time format	17:05; 19:30

b. Location Rule

Location entities are identified using location prefixes and preposition for places. For instance, "*Long Lamai*" is recognized as a Location since "Long" is a location prefix. In "*di Limbang*", "*Limbang*" is recognized as a Location because "*di*" is a locative preposition or <preposition><Proper Noun> -*di Limbang*.

c. Organization Rule

Organization is recognized by using the organization prefix or suffix. For example, "*kompeni Teras Kimia*" is an Organization because of the prefix "*kompeni*". "*Teras Kimia Sdn Bhd*" is an organization because of the suffix "Sdn Bhd" or <Organization prefix><Proper Noun><Organization suffix> - *kompeni Teras Kimia Sdn Bhd*. or <Proper Noun><Organization suffix> - *Teras Kimia Sdn Bhd* or <Organization prefix> <Proper Noun><Organization suffix> - *kompeni Teras Kimia Sdn Bhd.* In this case, there are 3 patterns of rules to identify the same entity. Thus, we use the most frequent used patterns. Besides that, we can identify organization by using the symbolic "()". For instance, Organization name (Abbreviation) - *Pengangkut Jalai Alun (JPJ).*

d. Date Rule

As for date, we can use the document Meta information. Meta means that the data in documents can be used directly most of times without further processing. For instance, news often starts with a location name plus date - *SIBU, 19 Jul.* Besides that, we can use the gazetteer lookup, lookup for the list of month. For instance, <Number><list of month> <Number> - *4 Disember 2004*.

e. Time Rule

As for time, we can identify time by using some trigger words such as 'pukul', 'pagi'(morning), 'tengah hari' (afternoon), 'lemai'(evening) . For instance, *pukul 6.45 pagi – pukul* [trigger words] *6.45* [Numeric] *pagi* [trigger words].

f. Percentage Rule

As for percentage, we can identify percentage by using the symbolic (%) or word (peratus). In Iban texts, they often use word compared to symbolic. For example*, 3.8 peratus.* Besides that, we also can use some specific word such a "niki" (climb). E.g. *niki 14.5 peratus (increase 14. 5%).*

g. Monetary Rule

As for monetary, we can monetary by using the symbolic of currency or currency unit. Thus, we need to build the list for symbolic currency and currency unit. Example of symbolic currency such as 'RM', '$' and etc; currency unit - rupiah, ringgit and etc.

4 Experiment Setup and Metrics

4.1 Experimental Setup

To evaluate NERSIL, we used an Iban annotated corpus made of 100 texts (29837 words, average 298 per texts), 22 lists of entities, and a set of Iban Jape rules (Table 4 and Table 5). The lists of entities constitute the Iban gazetteers.

Table 4: Number of Iban Jape Rules

Entity	Number of Jape rules
Person	14
Organization	5
Location	5
Time	4
Monetary	4
Date	3
Percentage	3

Table 5: Number of Jape Rules which Reuse and Created for Iban

Entity	No. of rules in ANNIE	No. of ANNIE rules that reuse	No. of created rules for Iban
Person	15	2	12
Organization	20	1	4
Location	7	1	4
Time	12	1	3
Monetary	2	1	3
Date	21	1	2
Percentage	2	1	2

GATE provides some evaluation metrics to assess the accuracy of the NER tool (Maynard, 2011). Recall measure the number of correctly identified item as a percentage of the total of correct item. Precision measure the number of correctly identified items as a percentage of items identified. F-measure is the average of the recall and precision.

$$\text{Recall:} \quad \frac{\text{Correct} + 0.5\ \text{Partial}}{\text{Correct} + \text{Missing} + 0.5\ \text{Partial}} \qquad (1)$$

$$\text{Precision:} \quad \frac{\text{Correct} + 0.5\ \text{Partial}}{\text{Correct} + \text{Spurious} + 0.5\ \text{Partial}} \qquad (2)$$

$$\text{F-Measure:} \quad \frac{(\beta\text{^}2 + 1)\, P * R}{(\beta\text{^}2) + P} \qquad (3)$$

In these formulae, 'correct' is understood as the number of correct recognition performed by the test system. 'Partial' is the number of partial correct recognition performed by the tested system. For instance, "Annah Rais" should be recognizing as a PERSON but the system recognized just "Annah" or "Rais" as a PERSON.

4.2 Experiment

To test the accuracy of NERSIL, we compared its results against human annotated texts and also the results of ANNIE on the same set of texts. It is expected that ANNIE will recognize only English NEs.

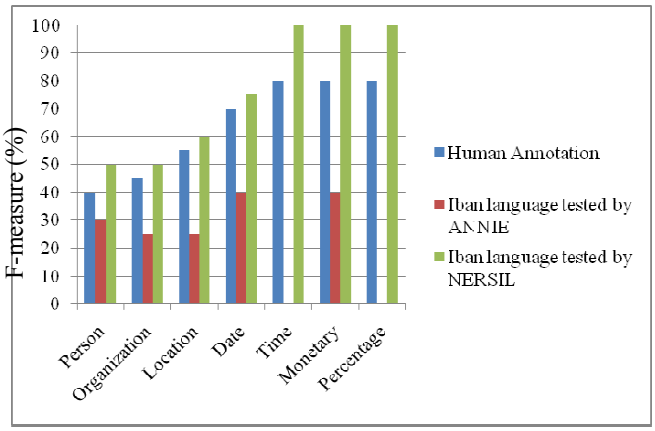

Figure 2: NE results on Iban texts

The overall results are shown in Figure 2. Human annotation achieves on F-measure of 64.3%, while Iban language tested by ANNIE achieves on F-measure of 22.8% and Iban language tested by NERSIL achieving on F-measure 76.4%. Based on these results, we noticed that ANNIE is also able to identify some NEs in Iban texts, which as we mentioned earlier are generally from English vocabulary. Therefore, we try to keep the default English gazetteers and add new named entities related to Malaysian locations to improve the performance of NERSIL. Besides, there are some common patterns we can find in English texts as well as in Iban texts such as date and monetary patterns. Hence, we can reuse the same JAPE rules for English and Iban. In the other hand, ANNIE is unable to identify NEs for time and percentage (Table 6), thus, we need to create specific rules for Iban texts.

Table 6: Difference between English language and Iban Language

Categories	English language	Iban language
Time	3:31:00 PM or 12:00 AM	Pukul 6 lemai or Pukul 10.30 malam
Percentage	10.0 %	10.0 peratus

Figure 3 shows the distribution of NEs in our Iban texts.

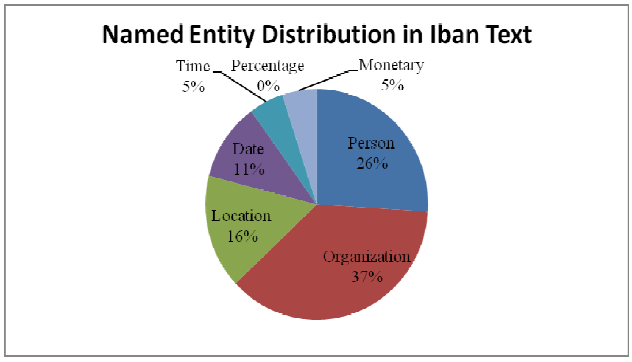

Figure 3: NEs Distribution in Iban Text

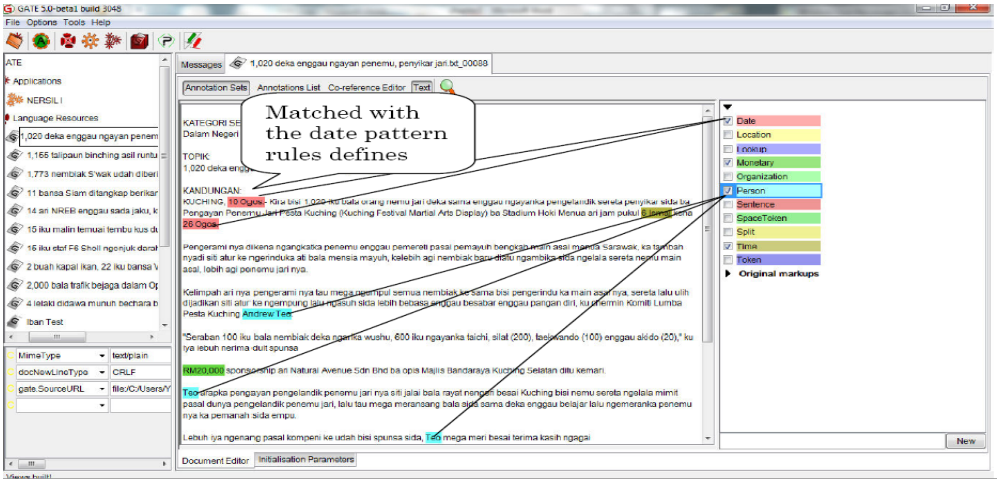

Figure 4: Architecture of Iban NE system (NERSIL)

5 Error Analysis

NERSIL is not perfect yet and some errors occurred during its testing.

- **Error in capital letters** – Some of the NEs was not capitalized. For example, chong ai ling. It should be Chong Ai Ling.
- **Spelling mistakes** – Some of the NEs was spelt incorrectly. For example, Cristopher Lee. The correct spelling is Christopher Lee. Both spelling mistakes and error in capital letters need another approach to solve it, for example, using a spelling checker for Iban language.
- **Incomplete gazetteers** – There are many NEs that occur in Iban texts and cannot be found in the current gazetteers especially Person, Location as well as Organization. Thus, this problem can be solved by improving the gazetteers.
- **Ambiguous words** – Homonymy is another source of errors. For example, Malaysia can be either a Location or an Organization. A Word Sense Disambiguation tool may solve this problem.
- **Missed tags by native speaker** – As our native speaker was not familiar with NEs and also GATE annotator, some tags were missing. Thus, the answer key annotations were inaccurate.

6 Conclusion and future work

In this project, we presented the development and testing of NERSIL, the first Iban NER tool. Based on our research, we can say that GATE is suitable for developing a NER tool for a new language. NERSIL was able to achieve 76.4% accuracy in identifying Person, Location, Organization, Date, Time, Percentage, and Monetary.

The contribution of this work is the creation of two language resources for Iban language that are NE rules and gazetteers. Moreover, we have adapted ANNIE to Iban. In addition, we have also helped in preserving an indigenous language.

While testing this Iban NER tool, few problems occurred. To overcome these problems, we propose, 1) Expanding the gazetteers, 2) Testing more Iban texts to discover more NEs, 3) Reducing negative effects on evaluation results – due to incomplete annotation of the test corpus, 4) Enhancing quality of transliterated names, 5) Using Iban texts with error free spelling, 6) Including all possible spelling variations used for names in Iban written texts.

References

Babych, B. and Hartley, A. 2003. *Improving Machine Translation Quality with Automatic Named Entity Recognition*. Proceedings of EACL-EAMT. Budapest. .

Budi, I. Bressan, S., Wahyudi, G., Hasibuna, Z.A. and Bazief, B.A.A. 2005. *Named Entity Recognition for the Indonesian Language: Combining Contextual, Morphological and Part-of-Speech Features into a Knowledge Engineering Approach.* Proceedings of 8th International Conference, DS 2005, Singapore, October 8-11, 2005

LDC. 2006. *Simple Named Entity Guidelines for Less Commonly Taught Languages, Version 6.5.* http://projects.ldc.upenn.edu/LCTL/Specifications/SimpleNamedEntityGuidelinesV6.5.pdf

Maynard, D., Valentin, T. and Hamish, C. 2003. *NE recognition without training data on a language you don't speak.* In Proceeding of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, pp. 33-40

Maynard, D. 2011. *Developing Language Processing Components with GATE,Version 6* http://gate.ac.uk/sale/tao/tao.pdf

Ng, E.L., Chin, B., Yeo, A. & Bali, R.M. 2010. *Identification of Closely-Related Indigenous Languages: An Orthographic Approach*. Intertional Journal on Asia Language Processing 20(2):43-61