Journal of ICT, 12, 2013, pp: 121–146

PROJECTING NAMED ENTITY TAGS FROM A RESOURCE RICH LANGUAGE TO A RESOURCE POOR LANGUAGE

[1]Norshuhani Zamin, [2]Alan Oxley and Zainab Abu Bakar[3]

[1, 2] *Faculty of Science and Information Technology,*
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia

[3]*Faculty of Computer and Mathematical Sciences,*
Universiti Teknologi MARA
40450 Shah Alam, Selangor, Malaysia

Corresponding author: norshuhani@petronas.com.my[1]

ABSTRACT

Named Entities (NE) are the prominent entities appearing in textual documents. Automatic classification of NE in a textual corpus is a vital process in Information Extraction and Information Retrieval research. Named Entity Recognition (NER) is the identification of words in text that correspond to a pre-defined taxonomy such as person, organization, location, date, time, etc. This article focuses on the person (PER), organization (ORG) and location (LOC) entities for a Malay journalistic corpus of terrorism. A projection algorithm, using the Dice Coefficient function and bigram scoring method with domain-specific rules, is suggested to map the NE information from the English corpus to the Malay corpus of terrorism. The English corpus is the translated version of the Malay corpus. Hence, these two corpora are treated as parallel corpora. The method computes the string similarity between the English words and the list of available lexemes in a pre-built lexicon that approximates the best NE mapping. The algorithm has been effectively evaluated using our own terrorism tagged corpus; it achieved satisfactory results in terms of precision, recall, and F-measure. An evaluation of the selected open source NER tool for English is also presented.

Keywords: Named entity recognition, information projection, bitext alignment, resource poor language, unsupervised learning, Malay terrorism corpus

INTRODUCTION

Named Entity Recognition (NER) is the fundamental task-oriented step in any Information Extraction (IE) application. It has been an essential task in various areas of Natural Language Processing (NLP) such as text summarization, document indexing, classification and translation. It has also served as a major research theme in the Message Understanding Conferences (MUC) and Conferences on Natural Language Learning (CoNLL) (Wahiba, 2011). In 1996, MUC was focusing on IE tasks where structured information is extracted from unstructured texts. MUC compiled defense-related news articles in MUC-3 and MUC-4 and Wall Street Journal articles in MUC-6. Tagged dataset corpora for IE tasks are also made available from these MUCs (Nadeau & Sekine, 2007). In NER, a system attempts to identify all mentions of proper names and quantities and categorize them into some predefined taxonomy of entities. MUC-7 defined the following Named Entity (NE) types which then were widely used in most NER research: person's names, geographic locations, organizations, dates, times, monetary amounts and percentages. Commonly, in any text corpus, proper names occur most frequently. In the MUC corpora alone, about 45-50% of the tags are organization tags, 12-32% are location tags while 23-39% are person tags (Feldman, & Sanger, 2007). In this article, 'person' (PER), which represents the name of a person, 'organization' (ORG), which represents the name of a company, association, etc., and 'location' (LOC), which represents the name of a geographical location such as a city, a country and also a building are being focused. Most research in NER has been dealing with un-annotated and unstructured blocks of text as in the string "John and Jane joined IBM in New York in 2006" to produce annotated text as in "(John/PER) and (Jane/PER) joined (IBM/ORG) in (New York/LOC) in (2006/DATE)".

NEs' characteristics

Different characteristics to determine the NEs are observed. An orthographic criterion is the main one as most NEs are identified by the capitalization. This rule applies to most of the alphabetic typed languages for proper nouns except for German which capitalizes all of its nouns. Some examples of English proper nouns are 'Bill Gates,' 'White House,' 'Sunset Boulevard' and 'Mitsubishi'. Proper nouns do not usually translate from one language to another. They often share spelling and sometimes are universal. As in the previous examples, all of the proper nouns remained unchanged in the Malay language except for White House which is translated into 'Rumah Putih.' NEs usually refer to a unique identity such as 'Barrack Obama,' who is the President of the United States

in 2012, and 'iPhone,' which is a model of mobile phone marketed by Apple. Although proper nouns carry reference because they specify an individual entity, they hold little semantic because they own limited attributes (Semenza, 1997). However, some words appearing before and after the proper nouns can be excellent indicators, such as 'The President,' 'Mr.,' 'Professor,' 'Limited,' 'Trading,' etc. These can be treated as entity rules for disambiguation. For example, a rule would help to correctly identify a person in 'Mr. Washington' and a location in 'Washington D.C'.

NER in different genres and domains

The impact of textual genre (journalistic, scientific, informal, etc.) and domain (sports, medical, political, business, gardening, etc.) plays an important role in any NER system's performance. Typically, NER systems perform well in a specific domain. There are NER systems designed for scientific and religious text (Maynard, Tablan, Ursu, Cunningham, & Wilks, 2001), email documents (Minkov, Wang, & Cohen, 2005; Jansche, & Abney, 2002; Gruhl, Nagarajan, Pieper, Robson, & Sheth, 2009), medical literature (Tanenblatt, Coden, & Sominsky, 2010; Han, & Ruonan, 2011) and newswire articles and web pages (McCallum, & Lee, 2003; Etzioni, 2005). It is a major challenge to generalize an NER system. Most NER systems perform reasonably well in their own domain. An experiment done on the performance of selected NER systems on various domains showed a drop of 20% to 40% of precision (P) and recall (R) (Poibeau, & Kosseim, 2001). The tested domains included the MUC-6 newswire articles, manual translation of phone conversation and technical emails.

NER systems that are trained on a single domain perform poorly out-of-domain. Some work has been established on open-domain NER to address this issue. An open-domain NER system is aimed at serving as a generic NER system that can produce accurate results in new domains. An open-domain NER system is able to identify the prominent entities in any scenario context without a priori knowledge (Nicolov, 2004). On the other hand, customizing existing NER annotators is a particularly challenging aspect of rule-based NER. It was recently reported (Chiticariu et al., 2010) that a significant amount of manual effort is required to perform the basic steps in building domain customization into an NER, which includes the identification of unambiguous semantic changes for the new domain, identification of the core annotator that should be modified and the development of the customization rules. This highly labor-intensive research does not receive much attention except in some literature (Wu et al., 2009; Giouli et al., 2006).

NER in different languages

NER research has been actively conducted in many languages for more than 20 years. Unfortunately, much attention was given to resource-rich languages such as English, Spanish and French (Pinnis, 2012). Resource-rich languages are the languages with almost a complete source of annotated data for NLP research use. These include the annotated corpora, engineered grammars, parsers, morphological analyzers, etc. They have been greatly exploited to produce many state-of-the-art NLP systems (Georgi, at. al., 2006). However, observations made on NER research for resource-poor languages have found that increased attention was given at the beginning of the year 2000. The most commonly presented are Chinese (Gao & Li, 2005; Wan et al., 2011), Japanese (Utsuro, & Sassano, 2000; Isozaki, 2001), Greek (Boutsis, Demitros, Giouli, Liakata, Papageorgiou, & Piperidis, 2000; Lucarelli, Vasilakos, & Androusopoulus, 2007) and Italian (Cucchiarelli, & Velardi, 2001; Federico, Nicola, & Vanessa, 2002).

Many other less common languages have been explored and experimented upon such as Latvian and Lithuanian (Pinnis, 2012), Barque (Whitelaw & Patrick, 2003), Bulgarian (Da Silva et al., 2004; Georgiev et. al, 2009), Cebuano (Maynard et al., 2003; May et al., 2003), Danish (Bick, 2004), Scandinavian (Johannessen et al., 2005), Romanian (Hamza et al., 2003), Swahili (Rushin et al., 2010) and Turkish (Metin, et al. 2012; Kucuk & Yazici, 2012). More recently, Arabic and Hindi have started to receive a lot of attention as reported by Huang (2005), Sujan et al. (2008), Ekbal, & Bandyopadhyay (2009) and Rajesh et al. (2011). On the other hand, limited NER research is to be found for the languages used in South East Asian (SEA) countries. There exists NER work for Thai (Chanlekha, & Kawtrakul, 2004; Tongtep, & Theeramunkong, 2011), Vietnamese (Tran et al., 2007; Nguyen et. al, 2010), Indonesian (Budi et al., 2005; Budi, & Bressan, 2007) and Filipino (Lim et al., 2007).

A multilingual NER is an NER system that is able to recognize named entities in a wide variety of languages. It is also referred to as a cross-lingual NER. Poibeau (2003) presented a language independent framework for detecting named entities in 13 different languages. The differences between the writing systems, morphologies and grammars have become a grand challenge in this nature of work. Poibeau's monolingual effort heavily relied on the existence of the annotated corpus for each language and a classical rule-based system. Ralf and Bruno (2007) attempted a multilingual NER for 10 different languages using a bilingual approach. The research proposed the use of language pair lexicons extracted from bilingual news wire articles. It is the language-

independent approach that enables entity detection for new languages to be plugged into the system effortlessly. Cross-lingual text analysis applications are made possible using existing multilingual linguistic resources such as thesauri, lexicons and gazetteers. This state-of-the-art approach has motivated us to undertake this research.

NER approaches

Hofmann (2001) described two schools of thought in NLP research. The first is the traditional linguistic community who believe that computerized learning is through linguistic theories and logic while the second is the statistically-oriented community who believe that a computer can learn from training examples such as document collections and corpora. This interesting phenomenon has made the NLP research, a 'never ending story'. Research in NER has been going on for more than 20 years. In the early years, hand-crafted rule-based algorithms were heavily explored. Today, modern systems most frequently use the machine learning approaches that involve purely statistical or probabilistic based techniques. The hand-crafted rule-based systems usually outperform the machine learning systems. Among the advantages of rule-based systems are a smaller storage requirement and easy domain extension using expert linguistic knowledge (Kim, & Woodland, 2000). However, this knowledge engineering approach is very expensive. Collecting relevant linguistic knowledge from experienced linguists is a laborious task (Zamin, Oxley, Bakar, & Farhan, 2012a). Some examples of rule-based NER systems are the identification of proper names in Greek financial texts using a hand-crafted lexical resource (Farmakiotou, Karkaletsis, Koutsias, Sigletos, Spyropoulos, & Stamatopoulos, 2000), one using a Constraint Grammar based parser for Danish (Bick, 2004) and an Arabic one for 10 different entities (Shaalan, & Raza, 2009). Gazetteer look-ups are often used as additional 'tools' to aid the rule-based NER systems (Zamin, & Oxley, 2011).

Currently, the leading approach to developing NER systems is supervised learning (Nadeau, & Sekine, 2007). It has been highly successful in solving many NLP tasks including NER. Supervised learning is an approach to automatically identify prominent entities from a collection of training examples. It is also able to infer learning rules from the data. However, supervised learning requires huge training examples before it is able to classify new/unseen data. In the absence of insufficient training data, semi-supervised and unsupervised learning are the other options but in the case of the unavailability of a training example, hand-crafted rule-based systems remain the preferred option. Supervised learning involves complex statistical techniques to examine the features of examples from a huge annotated corpus. An annotated corpus is a collection of labeled examples of potential input

paired with the corresponding correct output. A fully supervised NER system is able to predict named entities from a collection of pre-tagged documents. Supervised learning techniques include Maximum Entropy Models (Bender, Och & Ney, 2003), Conditional Random Fields (CRF) (McCallum & Li, 2003), Hidden Markov Models (Zu, & Sue, 2002), Decision Trees (Isozaki, 2001) and Support Vector Machines (Ekbal, & Bandyopadhyay, 2008).

Semi supervised learning typically takes a small amount of labeled data and a large amount of unlabeled data for training. It falls between the supervised and unsupervised learning methods. Semi supervised NER has been experimented with and is briefly discussed by Nadeau (2007). The thesis describes a semi supervised learning technique called 'bootstrapping' which requires little supervision. It takes a small set of labeled data referred as the 'seeds' to initiate the learning process. Some successful semi supervised NERs using the 'bootstrapping' technique are demonstrated by Sari, Hassan and Zamin (2009), Pasca, Lin, Bigham, Lifchits & Jain (2006) and Heng & Grishman (2006). Unsupervised learning is a method where the machine learns from unlabeled data. This learning method commonly deals with problems without a target output or a reward signal in order to evaluate potential solutions. The challenge is how to get the machine to successfully induce learning without there being any response from the environment. Ghahramani (2004) proposed a solution using a "formal framework based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, effciently communicating the inputs to another machine, etc." Among the methods used in unsupervised learning are Artificial Neural Networks, Cluster Analysis, Self-Organizing Maps and Expectation Maximization. Some NER work includes that of Nadeau (2006), Alfonseca, & Manandhar (2002) and Ciaramita, Gamgemi, Ratsch, Saric & Rojas (2008). There are also hybrid solutions as shown in the work of Srihari (2000), Wu, Zhao, & Xu (2003) and Szarvas, Farkas & Kocsor (2006).

Poorly resourced languages

Singh (2008) discussed some issues on linguists' perspectives of less resourced, or less privileged, languages. He commented and presented noteworthy facts on the research at the Workshop on NLP for Less Privileged Languages. Singh pointed out that there exist machine translation for Basque, a predictive text input system for Sinhala, a finite state solution for reduplication in Kinyarwanda and a part-of-speech (POS) tagger for Manipuri. A comparative study between monolingual and cross-lingual bootstrapping, reported in Chen & Ji (2009), provides a better view on the significance of cross-lingual research for poorly resourced languages. Additionally, Singh et al. (2008) has studied the problem to estimate the resource adaptation cost from a resource-

rich to a resource-poor language and found that creating resources for most NLP work is extremely complicated for any language. It is even harder for under-privileged languages that are less computerized. They propose a state-of-the-art solution to this by "adapting the resources of a linguistically close resource-rich language." The central idea is to project the annotated resources of a resource-rich language to the second language. The advantages are less labor, less expense and a faster turnaround time in producing a solution. It is also referred to as bitext alignment, word alignment of parallel/bilingual corpora, text mapping or resource projection in most of the literature. Ma (2010) addressed two main steps in bilingual NER: 1) identification of the entities in both halves of the bilingual corpus and 2) alignment of the entities across two languages.

Cross-language annotation projection research for NER was pioneered by Yarowsky, Ngai & Wicentowski (2001). They demonstrated statistical algorithms to perform basic tasks in text analysis, which are POS tagging, base noun phrase bracketing, NER and inflectional morphological analysis. Established English corpora were used to align the annotations to four different languages - French, Chinese, Czech and Spanish. In its NER module, the MUC-6 corpus is used to tag the 'person,' 'location' and 'organization' entities in French-English Canadian Hansards data. It does so with an accuracy of 64%. The underlying algorithm is based on the projection/bridge-based similarity measure. More recently, Ma (2010) introduced co-training on unlabeled bilingual data to adapt existing NE taggers to new domains and thus improve the current state-of-the-art of NE taggers. The co-training algorithm iteratively selects new training instances from unlabeled text to create annotated text. The work is demonstrated on a Chinese-English bilingual corpus available from Doddington et al. (2005). This baseline work is then used to tag named entities in another Chinese-English bilingual corpus and comparable text collections. However, to improve accuracy, they employed several techniques: 1) Pinyin Mapping – to perform Chinese into English transliteration, 2) Dictionary lookup – a lexicon of possible Chinese and English translations, 3) Transliteration Model – an existing Chinese to English transliteration for names and 4) Google Translation – the translation tools to translate Chinese texts into equivalent English. Other related projective NER works for different languages are 'person name identification' using 'crowd sourcing judgments' for 21 different languages (Mayfield et al., 2011), and a projection from an Arabic-English hand-aligned parallel corpus that was previously tagged with a baseline NER system known as BASE (Benajiba, Zitouni, Diab & Rosso, 2010). To the best of our knowledge, no literature exists on projection of the Malay language. This paper is the first work of its kind, a method which is new to the Malay language and possibly a noteworthy contribution to the Malay linguistic community.

METHOD

In this paper, the idea of using the resources available in a resource-rich language i.e. English is investigated, as an effective way to identify PER, ORG and LOC entities in Malay texts. The Malay language is spoken by 300 million people in SEA covering the countries of Malaysia, Brunei, Singapore and Indonesia (El-Imam, & Don, 2005). As a case study, Malay journalistic texts on terrorism are used to create the first corpus of Malay terrorism. The unavailability of such a corpus has initiated this study. To begin with, 25 news articles describing Indonesian terrorism, published by various news agencies and written in Malay, were used to test our proposed approach. In this work, a novel approach is devised. (It so happens that our selected approach can be used in other contexts than the one described here.) A simple, but an efficient basic approach that uses small, parallel, annotated corpora is proposed to automatically infer named entities in a Malay journalistic corpus of terrorism. The framework of our NE Tagger consists of several modules that are: a Malay-English Word Aligner, an NE Extractor and a Noise Filterer, as shown in Figure 1. The aim of this work is to project PER, ORG and LOC entities from English to Malay via a statistical word aligner algorithm to automatically create the NER Malay Corpus of Indonesian Terrorism. The UIUC NE tagger (Ratinov, & Roth, 2009) is employed. It is an off-the-shelf NER tagger and is used to tag our parallel English corpus.

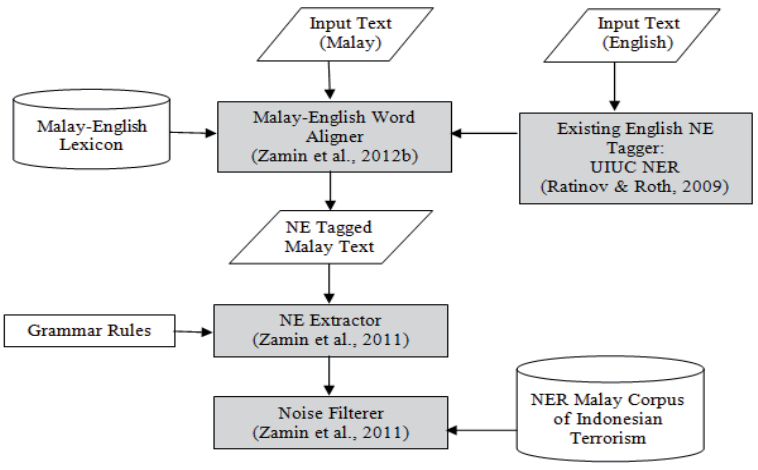

Figure 1. Framework of Malay NE Tagger.

Each of the elements and processes in the framework is described in the following sub-sections.

Input Texts

The study involves all sorts of terrorism related cases in SEA including those involving radicalism and violence. Among all the countries in SEA, in Indonesia and Thailand the number of terrorism incidents has increased substantially. Western intelligence is fingering Indonesia as a base of terrorism after a sequence of violent attacks by the largest clandestine terrorist network known as the Jemaah Islamiyah (JI). A total of 300 articles are collected and among these are 187 articles related to JI. Due to the prevalence of terrorist activities in Indonesia, 25 Malay journalistic articles reporting on Indonesian terrorism are selected. Collectively they consist of 263 sentences, or 5413 words. These were used to evaluate the framework. All the articles were digitized and translated into English using the Google Translate[1]. Every translation was validated by a human in order to increase correctness. The method does not use aligned bilingual corpora as in Yarowsky, Ngai & Wicentowski (2001) as such a resource is not available for the research domain. Both the Malay text and its translated version form the parallel corpora and serve as the input to the proposed system.

Malay-English Lexicon

A Malay-English lexicon, a dictionary look-up containing all possible lexemes in English for a given Malay word is proposed. Lexemes help to avoid lemmatizing inflectional Malay words. Lemmatization is expensive work. It involves deep morphological analysis and a complex stemming algorithm depending on the structure of a language. The task of lemmatizing is to identify the lemma (base or root word) of a given word. For example, 'walk' is the lemma of 'walking' while 'walks', 'walked' and 'walking' are the lexemes for 'walk'. Some lexemes/lemmas can be easily identified by a rule-based lemmatizer but some need a gazetteer (list look-up) such as 'good,' the lemma for 'better' and 'best.' Figure 2 shows sample lexicon entries for the Malay word 'dakwa' (accuse). The lexicon is created using the free Online Dictionary Malay & English[2].

dakwa = accuse, claim	mendakwa = accuse, claim
berdakwa = litigate	mendakwakan = indict, bring to court
dakwaan = accusation, charge	pendakwaan = accusation, charge

Figure 2. Lexicon entries for the Malay word 'dakwa' (accuse).

[1] *http://translate.google.com.my*

[2] *http://kamus.lamanmini.com/index.php*

Existing English NE tagger

A comparative study on the performance of three open source NER Taggers over our English terrorism corpus is conducted. The three selected state-of-the-art taggers are Stanford, LingPipe and UIUC. The recognition of PER, ORG and LOC entities over 25 articles is tested. The NER system that returned the highest accuracy and was closer to human annotations is chosen as our 'English resource' for the bitext projection experiments. The Stanford NER Tagger (Finkel, Grenager & Manning, 2005) is a supervised tagger. It was developed using the CRF method and trained on the CoNLL03, MUC-6, MUC-7 and the Automatic Content Extraction (ACE) NE corpora. CoNLL, MUC and ACE are the significant bodies in computational linguistics whose efforts include conducting competitions, conferences and producing benchmark datasets and results to support NLP research. For example, the CoNLL03 NE data consists of eight files covering the English and German languages. The English data is a collection from the Reuters Corpus[3] while the German data is a collection from the ECI Multilingual Text Corpus. The MUC-6 and MUC-7 corpora are compilations of American newswire texts on management changes and satellite launch reports, respectively. LingPipe (Alias-1, 2008) is a Java API which was introduced by Baldwin from the University of Pennsylvania and Carpenter from the University of Edinburgh. It is made available under licensing terms that range from free to perpetual server licenses. LingPipe is a computational linguistics toolkit for text processing tasks including NER. The LingPipe NER Tagger was trained on English corpora of multiple genres and domains including MUC-6 (news), GeneTag (genes) and GENIA (genomics). UIUC NE Tagger (Ratinov, & Roth, 2009) was developed by the Cognitive Computation Group of the University of Illinois at Urbana Champaign. UIUC is an English NER tagger using gazetteers extracted from Wikipedia, word class models derived from unlabeled text and expressive non-local features. This supervised learning tagger was trained on the CoNNL03 and MUC-7 datasets which have been previously annotated with at least four 'classic' entity type sets (people / organizations / locations / miscellaneous). The performance of these three NER taggers is compared because, to the best of our knowledge, they are the best publicly available systems and were trained on almost the same data.

However, none of these taggers was trained on a journalistic corpus of terrorism. Therefore, an experiment was conducted to test the performance of each tagger over the new domain. The classical evaluation metrics for text processing is used, that is, P, R and the F1-Score (F1). Briefly, P is the proportion of names proposed by the system which are true names while R

[3] *http://www.reuters.com/researchandstandards/*

is the proportion of true names which are actually identified. These metrics are often combined and referred to as F1. Hence, F1 is a weighted harmonic between P and R. The metrics' expressions are as follows: P = correct / (correct + wrong), R = correct / (correct + missed) and F1 = 2PR / (P + R). The results are presented in Table 1.

Table 1

A Comparative Study between Stanford, LingPipe and UIUC NE Tagger Using English Terrorism Corpus

NE Tagger	PER			ORG			LOC		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Stanford	0.76	0.59	0.66	0.59	0.42	0.49	0.89	0.79	0.84
LingPipe	0.67	0.56	0.61	0.62	0.61	0.62	0.73	0.72	0.72
UIUC	0.97	0.77	0.86	0.93	0.84	0.88	0.86	0.82	0.84

The best results from this experiment were achieved by the UIUC NER Tagger. The key to its success is the Wikipedia-based gazetteers. The NE list is produced by extracting from Wikipedia[4]. Wikipedia is an open, collaborative encyclopedia with a number of attractive properties. New entities are manually and constantly added by Wikipedia's collaborators. Wikipedia is able to redirect pages and map several variations of spelling of the same name to one canonical entry.

Word aligner

The Malay-English Word Aligner algorithm is a potential method for aligning two languages of dissimilar patterns and structures. The algorithm combines a bigram scoring method, to assess the similarity of two strings, and a Dice Coefficient function (Dice, 1945), to measure the string closeness of two different texts. Bigram scoring is applied to all the possible English lexemes extracted from the lexicon for a given Malay word. The sequence of graphemes in each lexeme is compared with all the English words in the English text. Bigram scoring is used to pick the English lexeme with the highest similarity. This is a simple and faster solution to building a complex morphological analyser for Malay. Consequently, our unsupervised approach does not require any hand-tagged labelled data. This research is inspired by the work of Dien (2001, 2005). However, some limitations are observed in these works. The morpheme similarity measure requires a morphological

[4] *http:// www.wikipedia.org*

analyzer for the Vietnamese language and the effort to pre-align the bilingual corpus is a laborious and expensive approach. There is still, technically, a lot of room for improvement. For this reason, the following bigram score and Dice Coefficient function is proposed to map the English word and the Malay word, thus projecting the corresponding NE tag from the English word to its Malay translation:

$$Sim(d_i, E_j) = \frac{2 \times N_{d_i \cap E_j}}{N_{d_i} + N_{E_j}} \quad (1)$$

where is the number of bigrams common to both an English lexeme in the lexicon (d_i) and an English word in the English corpus (E_j). is the number of bigrams found in d_i, and is the number of bigrams found in E_j. Figure 3 shows an example of bigram pair-wise matching for the word 'the' against the lexemes 'unbelievable' and 'unreliable.' The technical details of this algorithm are given by Zamin, Oxley, Abu Bakar & Farhan (2012b) with worked examples.

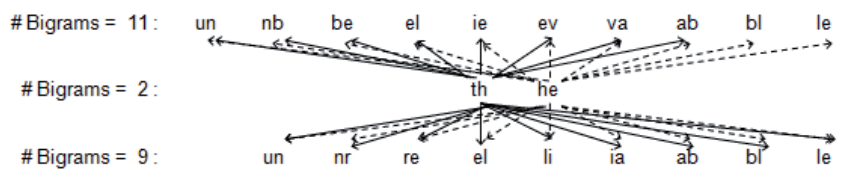

Figure 3. Bigram pair-wise matching of English words.

All the proper names appearing in the corpus are also stored as they are in our lexicon. No prior categorization, hand-tagged annotation or pre-alignment of words is done. The NE tags are directly projected from the English source, which was previously tagged by the UIUC NER Tagger. Hence, there is no ambiguity problem as they are already resolved by the open tagger. In contrast to look-up based NER systems, they commonly suffer from limitations of coverage and ambiguity. For example, the word 'Washington' in 'President Washington' will not be recognized as a person if the entity 'Washington' was previously defined as a location in the list. However, there is a potential way to integrate grammar rules for further proper name disambiguation, as is described in the next section.

Grammar rules

Knowledge-based approaches, such as rules and gazetteers for NE tagging, were much explored in the early years. The explicit resources are usually handcrafted by experienced language experts and do not require any training

data. This laborious method commonly performs better than the machine learning approaches. Hence, NE tagging would be well served by using a hybrid methodology that incorporates rule-based and machine learning approaches, where appropriate, as is demonstrated in Kim & Compton (2012). In this paper, the grammar rules are the common patterns to identify proper names in an English corpus. Some rules for regular English expressions from Feldman & Sanger (2007) are adopted and combined these with the manually created rules through observation of the translated terrorism corpus and by using the projected POS tag for proper names. Table 4 shows several grammar rules used to identify a person's name (PER) and locations (LOC).

Table 4

Examples of Grammar Rules

Rule	***Variable / Description***	***Example***
@Honorific1/nnp @ Honorific2/nnp PER1/nnp (PER2/nnp)(PER3/nnp)	*@Honorific* can be in multiple forms: Major General, Brigadier General, Datuk Seri, Prime Minister	Brigadier General **Sulistiyo Ishak**
PER1/nnp (PER2/nnp) @ Verb/vbd	*@Verb* is a common verb that is strongly associated with people: said, met, walked, etc.	**Danuri** said the raid was targeting Jemaah Islamiyah.
@FirstNames/nnp PER1/nnp (PER2/nnp) (PER3/nnp)	@FirstNames is a common first name collected from the corpus.	Susilo as in **Susilo** Bambang Yudhoyono; Noordin as in **Noordin** Mat Top
@Prep/in LOC1/nnp (LOC2/nnp) (LOC3/nnp)	@Prep is a preposition commonly associated with locations: on, in, at	Three men were found guilty of conducting bomb attacks on night clubs in **Kuta, Bali.**

(NOTE: nnp = Proper Noun Tag, in = Preposition Tag, vbd = Verb Tag)

Noise filterer

A dictionary matching scheme is often vulnerable to false positives. A false positive is a case where a proper name identified by the Entity Extractor is in fact a non-name and can be considered as noise. False positives often degrade such a system's accuracy. Hence, a Noise Filterer module is added to the framework to remove the unwanted names by simply eliminating low-confidence predictions (Zamin et al., 2011). There are two metrics used in this module, as introduced by Minkov, Wang & Cohen (2005) - Predicted Frequency (PF) and Inverse Document Frequency (IDF). The PF metric

estimates the degree to which a word appears to be used consistently as a proper name throughout the corpus.

$$PF(w) = \frac{cpf(w)}{ctf(w)} \tag{2}$$

where cpf(w) is the number of times that a word w is identified as a name and ctf(w) is the number of times it appears in the entire test corpus. The IDF metric is calculated as follows:

$$IDF(w) = \frac{\log(\frac{N+0.5}{df(w)})}{\log(N+1)} \tag{3}$$

Where df(w) is the number of articles that contain the word w and N is the total number of articles in the corpus. PFIDF is a measure which combines these two metrics multiplicatively, giving a single probability of a word being a name and showing how common it is in the entire corpus; the measure is as follows:

$$PFIDF(w) = PF(w) \times IDF(w) \tag{4}$$

A word with low PFIDF score is considered ambiguous in the corpus and is excluded from it.

NE Extractor

The NE Extractor is a module that extracts person, organization and location entities based on the projected NE tags (PER, ORG and LOC) and the grammar rules, as is illustrated in Figure 4.

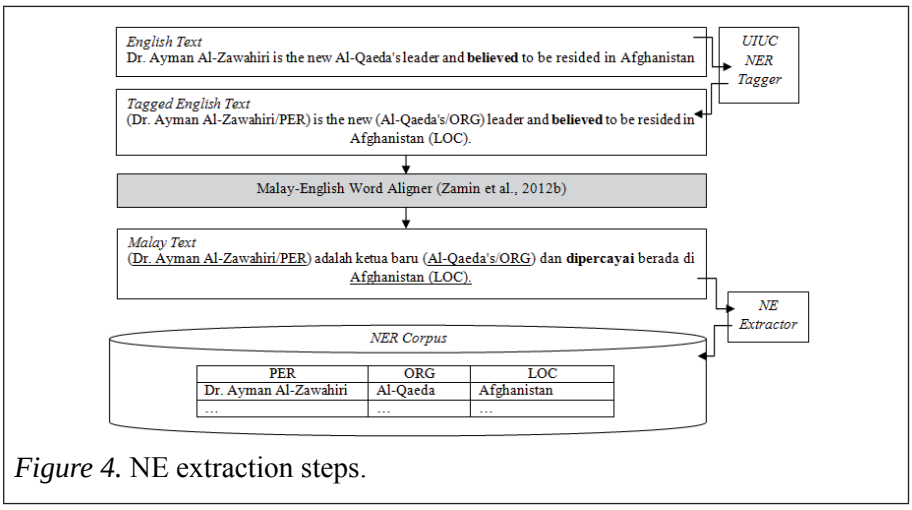

Figure 4. NE extraction steps.

NER Malay corpus of Indonesian terrorism

The output of the system is a structured set of entity types as shown in Table 2.

Table 2

Examples of the Extracted Named Entities in a Structured Presentation

Article Information	*PER*	*ORG*	*LOC*
ID: IND001 Source: Utusan Malaysia Date: 02032010 Title: Indonesia dakwa 4 lelaki pengganas. (#Sentences = 10, #Words = 184)	Esa Permadi Sulistiyo Ishak	Jemaah Islamiyah Associated Press Jakarta Globe	Jawa Banda Aceh
ID: IND002 Source: Berita Harian Date: 11032010 Title: Susilo janji buru pengganas. Sub-title: Presiden Indonesia sah Dulmatin ditembak mati. (#Sentences = 11, #Words = 257)	Susilo Bambang Yudhoyono Susilo Dulmatin	Jemaah Islamiyah	Indonesia Pamulang Jakarta Filipina Australia Bali
ID: IND003 Source: Berita Harian Date: 12032010 Title: Polis temui alat peledak, senjata api. Sub-title: Serbuan bongkar rancangan letup kafe Internet di Jakarta. (#Sentences = 7, #Words = 174)	Bambang Hendarso Danuri Danuri Dulmatin Susilo Bambang Yudhoyono	Jemaah Islamiyah Al-Qaeda	Indonesia Bali

RESULTS

Performance of the system was measured, with P, R and F1 as the measures, for 5413 words extracted from 25 Malay journalistic articles describing Indonesian terrorism. However, due to the non-existence of an NER-annotated corpus for Malay, particularly in the terrorism domain, a small golden corpus for Malay with approximately 5000 words is manually constructed, in order for us to carry out our evaluation. The performance of the system is evaluated based on two different experiments. The first experiment was for testing the system performance with the word aligner algorithm alone, i.e. the machine learning approach, whilst the second experiment was for testing the hybrid method, i.e. the combination of the word aligner algorithm with the grammar rules. The experiments tested the PER, ORG and LOC entities only. Table 3 presents the performance of our algorithm using both the unsupervised (Word Aligner) and the hybrid (Word Aligner + Grammar Rules) methods on the datasets.

Table 3

The Results of Projecting NE Tags between Parallel Terrorism Corpora - from English To Malay

Method	*Correct Tags*	*Incorrect Tags*	*Missed Tags*	*P*	*R*	*F1*
Word Aligner	400	84	74	0.83	0.72	0.77
Word Aligner + Grammar Rules	494	34	35	0.94	0.88	0.90

An NE is considered correct only if it is an exact match to the corresponding entity in the hand-tagged dataset. Overall, the Word Aligner alone achieved a performance of 77% but this increased to 90% when grammar rules were added. That is, the hybrid method improved F1 by 13%. It shows that the attempt to integrate rules to help identify prominent entities that could be missed by the statistical algorithm is a worthwhile effort. It is observed from the experimental results that more than 25% of the recognition mistakes were due to different spellings used in the lexicon and news articles, no matching of a lexeme in the lexicon, and the errors produced by the UIUC NER Tagger itself. Incorrect spelling of proper nouns in the lexicon results in misidentification of the entity, for example 'Banda Aceh vs. Banda Acheh' and 'Jemaah Islamiyah vs. Jemaah Islamiah.' A variation in spelling is found between different news agencies. Besides, the performance can be increased by adding more detailed rules with fixed variable strings, such as 'hotel,' 'province,' 'north of,' 'district,' 'new,' 'agency,' etc. Additionally, mapping multiple words of different sizes such as 'Amerika Syarikat' and 'United States of America' (being its equivalent translation in English) contributes to an increase in error rate. Finally, incorrect tagging of capitalized nouns and missed tagging of entities by the UIUC NER Tagger also affect the output of the system. Table 4 presents some erroneous tagging produced by the UIUC NER Tagger using our dataset. Words given the wrong tag are underlined.

Table 4

UIUC's Tagging Errors

Error Type	*Example*
Missed Tagging	*Example 1:* `MISC Southeast Asian) extremist group inspired by al-Qaeda and blamed for several attacks in (LOC Indonesia).`

(continued)

Error Type	*Example*
	Example 2: They are conducting military training in remote areas of the forests of (LOC Aceh), probably because (LOC Aceh) is now peaceful," he said, as quoted by the agency of Antara news today. *Example 1:* More than 100 heavy-armed police officers took part in the raid just before midnight last night in the forest areas of the (MISC Aceh Besar) district, about 70 kilometers north of (PER Banda Aceh).
Wrong Tagging	*Example 2:* Police made the arrests after fighting for an hour late yesterday in the mountainous area of (PER Jalin), said the Provincial Police Chief, Major General (PER Aditya Warman).

CONCLUSION

NER is a knowledge-intensive task. It is an important component in much NLP research, including IE and IR. Projecting explicit linguistic tags from another language via parallel corpora has been widely used in NLP tasks and has proven to contribute significantly to achieving better performance. The proposed Malay-English Word Aligner algorithm provides a successful bridge for aligning complex inflected word forms and projecting information in two dissimilar and highly irregular surfaces. Thus, our research contributions can be summarized as follows:

- A hybrid NER Tagger for a Malay journalistic corpus of Indonesian terrorism that employs an unsupervised learning method has been developed - a statistical technique using the classical Dice Coefficient function with bigram scoring and a rule-based method, and a set of knowledge engineered rules. Usage was made of a Malay-English lexicon as an add-on to overcome the costly lemmatization effort needed for the Malay language.
- Three state-of-the-art NER Taggers, i.e. the Stanford, the LingPipe and the UIUC have been evaluated on our English dataset. The performance was decisively low for both the Stanford and the LingPipe NER Taggers. The most probable reason for the poor outcomes are the different genre and domain used, ones that were never tested previously on these taggers. However, the use of a gazetteer extracted from Wikipedia has shown an increased performance in the UIUC NER Tagger. Hence, UIUC was chosen as the NER resource for English.

- A module to filter unwanted names, those with low confidence, so as to increase the tagger's performance rate has been integrated to the framework.
- How the hybrid approach handles the projection of linguistic tags for both POS tagging (Zamin et al., 2012a; Zamin et al., 2012b) and NER tagging at a fairly accurate rate have been successfully demonstrated.

REFERENCES

Lucarelli, G., Vasilakos, V., & Androutsopoulos, I. (2007). *Named entity recognition in Greek texts*.

McCallum, A. & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the Natural Language Learning,* 188-191.

Alfonseca, E., & Manandhar, S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. *Proceedings of the International Conference on General WordNet,* 34-43.

Alias-I. (2008). LingPipe 4.1.0. Retrieved from http://alias-i.com/lingpipe. *Association of Computational Linguistic,* 31, 531-574. doi: 10.1162/089120105775299177.

Benajiba, Y., Zitouni, I., Diab, M., & Rosso, P. (2010). Arabic named entity recognition: using features extracted from noisy data. *Proceedings of the Association for Computational Linguistics Conference Short Papers,* 281-285.

Bender, O., Och, F. J., & Ney, H. (2003). Maximum entropy models for named entity recognition. *Proceedings of the Natural Language Learning,* 148-151.

Bick, E. (2004). A named entity recognizer for Danish. *Proceedings of the Language Resources and Evaluation*, 305-308.

Boutsis, S., Demitros, L., Giouli, V., Liakata, M., Papageorgiou, H., Piperidis, S. (2000). A system for recognition of named entities in Greek. *Proceedings of the Natural Language Processing,* 424-436

Budi, I., & Bressan, S. (2007). Application of association rules mining to named entity recognition and co-reference resolution for the Indonesian language. *International Journal of BI and DM*, *2*, 426-446.

Budi, I., Bressan, S., Wahyudi, G., Hasibuan, Z., & Nazief, B. (2005). Named entity recognition for the Indonesian language: combining contextual, morphological and part-of-speech features into a knowledge engineering approach. *Discovery Science*, 57-69. Springer Berlin/Heidelberg.

Chanlekha, H., & Kawtrakul, A. (2004). *Thai named entity extraction by incorporating maximum entropy model with simple heuristic information. Proceedings of the International Joint Conference in Natural Language Processing.*

Chen, Z., & Ji, H. (2009). Can one language bootstrap the other: A case study on event extraction. *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing,* 66-74.

Ciaramita, M., Gangemi, A., Ratsch, E., Šarić, J., & Rojas, I. (2008). Unsupervised learning of semantic relations for molecular biology ontologies. *Proceeding of the Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, 91-104.

Cucchiarelli, A., & Velardi, P. (2001). Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, *27*, 123-131.

Da Silva, J. F., Kozareva, Z., & Lopes, G.P. (2004). Cluster analysis and classification of named entities. *Proceedings of the Language Resources and Evaluation,* 321-324. doi: 10.1.1.99.4830.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*, 297-302.

Dien, D. I. N. H. (2001). *Building an English-Vietnamese bilingual corpus* (Unpublished master's thesis in comparative linguistics). University of Social Sciences and Humanity of Ho Chi Minh City, Vietnam.

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.

Doddington, G, Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. & Weischedel, R. (2004). *Automatic Content Extraction (ACE) program - task definitions and performance measures. Proceedings of the Language Resources and Evaluation*.

Ekbal, A., & Bandyopadhyay, S. (2008,). Bengali named entity recognition using Support Vector Machine. *Proceedings of the Workshop on NER for South and South East Asian Languages and International Joint Conference on Natural Language Processing,* 51-58.

Ekbal, A., & Bandyopadhyay, S. (2009). A Conditional Random Field approach for named entity recognition in Bengali and Hindi. *Linguistic Issues in Language Technology*, *2*, 1-44.

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.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D., & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, *165*, 91–134.

Evans, R. (2003). A framework for named entity recognition in the open domain. *Proceedings of the Recent Advances in Natural Language Processing,* 137-144, doi: 10.1.1.105.7150.

Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Rule-based named entity recognition for Greek financial texts. *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries,* 75-78.

Federico, M., Nicola, B., & Vanessa, S. (2002). Bootstrapping entity recognition for Italian broadcast news. *Proceedings of the Empirical Methods in Natural Language Processing,* 296-303. doi: 10.1.1.11.3083.

Feldman, R., & Sanger, J. (2007). *Text mining handbook*. Cambridge: Cambridge University Press.

Finkel, J.R., Grenager T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the Annual Meeting of the Association for Computational Linguistics,* 363-370.

Gao, J., Lee, M. (2005). Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, *31*, 531-574.

Georgi, R., Xia, F., & Lewis, W. (2012). Measuring the divergence of dependency structures cross-linguistically to improve syntactic projection algorithms. *Proceedings of the Language Resources and Evaluation,* 771-778.

Georgiev, G., Nakov, P., & Ganchev, K. (2009). Feature-rich named entity recognition for Bulgarian using Conditional Random Fields. *Proceedings of the Recent Advances in Natural Language Processing,* 113-117.

Ghahramani, Z. (2004). Unsupervised learning. *Advanced Lectures on Machine Learning,* 72-112.

Giouli, V., Konstandinidis, A., Desypri, E., & Papageorgiou, Harris. (2006). Multi-domain multi-lingual named entity recognition: Revisiting and rounding the resources issue. *Proceedings of the Associations for Computational Linguistics,* 59-64.

Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., & Sheth., A. (2009). Context and domain knowledge enhanced entity spotting in informal text. *Proceedings of the Semantic Web Conference*, 260-276. doi: 10.1007/978-3-642-04930-9_17.

Hamza, O., Bontcheva, K., Maynard, D., Tablan, V., & Cunningham, H. (2003). Named entity recognition in Romanian. *Technical Report, Department of Computer Science.* University of Sheffield.

Han, X., & Ruonan, R. (2011). The method of medical named entity recognition based on semantic model and improved SVM-KNN algorithm. *Proceedings of the Semantic Knowledge and Grid,* 21-27.

Heng, J., & Grishman, R. (2006). Data selection for semi supervised learning for named tagging. *Proceedings of the Workshop on Information Extraction Beyond The Document*, 48-55. Association for Computational Linguistics.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177-196.

Huang, F. (2005). *Multilingual named entity extraction and translation from texts and speech* (Unpublished doctoral dissertation). Pittsburgh: Carnegie Mellon University.

Isozaki, H. (2001). Japanese named entity recognition based on a simple rule generator and decision tree learning. *Proceedings of the Annual Meeting on Association of Computational Linguistics,* 314-321.

Jansche, M., & Abney, S. (2002). Information extraction from voicemail transcripts. *Proceedings of the Empirical Methods in Natural Language Processing,* 10, 320-327. doi: 10.3115/1118693.1118734.

Johannessen, J. B., Hagen, K., Haaland, A., Jonsdottir, A.B., Noklestad, A., Kokkinakis, D., & Meurer, P. (2005). Named entity recognition for the mainland Scandinavian languages. *Literary and Linguistics Computing*, *20*, 91-102.

Karaa, W.B.A. (2011). Named entity recognition using web document corpus. *Managing Information Technology, 3*. doi: 10.5121/ijmit.2011.3104.

Kim, J.H., & Woodland, P.C. (2000). Rule-based named entity recognition. *Technical Report CUED/F-INFENG/TR.385.* Cambridge University.

Kim, M., & Compton, P. (2012). Improving the performance of a named entity recognition system with knowledge acquisition. *Knowledge Engineering and Knowledge Management*, 97-113.

Kucuk, D., & Yazici, A. (2012). A hybrid named entity recognizer for Turkish. *Expert Systems with Applications*, *39*, 2733 – 2742.

Lim, N. R., New, J. C., Ngo, M. A., Sy, M., & Lim, N. R. (2007). A named-entity recognizer for Filipino texts. *Proceedings of the National Natural Language Processing Research Symposium,* 20-25.

Lucarelli, G., Vasilakos , X., & Androutsopoulos, I. (2007). Named entity recognition in Greek with an ensemble of SVMs and active learning, *Artificial Intelligence Tools Journal*, *16*, 1015-1045.

Ma, X. (2010). Toward a named entity aligned bilingual corpus. *Proceedings of the Language Resources and Evaluation*.

Mayfield, J., Lawrie, D., McNamee, P., & Oard, D. (2011). Building a cross-language entity linking collection in twenty-one languages. *Multilingual and Multimodal Information Access Evaluation*, 6941, 3-13. Springer-Verlag.

Maynard, D., Tablan, V., & Cunningham, H. (2003). NE recognition without training data on a language you don't speak. *Proceedings of the ACL Workshop on Multilingual and Mixed-Language Named Entity Recognition: Combining Statistical and Symbolic Models,* 33-40. Association for Computational Linguistics.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., & Wilks, Y. (2001). Named entity recognition from diverse text types. *Proceedings of the Recent Advances in Natural Languages Processing*, 257-274. doi: 10.1.1.18.7395.

McCallum, A., & Li, W. (2003). Early results for named entity recognition with Conditional Random Felds, feature induction and web-enhanced lexicons. *Proceedings of the Computational Natural Language Learning, 4*, 188-191. doi: 10.3115/1119176.1119206.

Metin, K.S., Kisla, T., & Bahar, K. (2012). Named entity recognition in Turkish using association measures. *Advanced Computing: An International Journal*, *3*, 43-49. doi: 10.5121/aclj.2012.3406.

Minkov, E., Wang, R., & Cohen, W. (2005). Extracting personal names from emails: applying named entity recognition to informal text. *Proceedings of the Human Language Technology and Conference on Empirical Methods in Natural Language Processing*, 443-450. doi:10.3115/1220575.1220631.

Nadeau, D. (2007*). Semi-supervised named entity recognition: Learning to recognize 100 entity types with little supervision* (Unpublished doctoral dissertation). University of Ottawa.

Nadeau, D., & Satoshi, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigaciones*, 30, 3-26. doi:10.1075/li.30.1.03nad.

Nadeau, D., Turney, P., & Matwin, S. (2006). Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity. *Advances in Artificial Intelligence: Conference of the Canadian Society for Computational Studies of Intelligence*, *19*, 266-277. Springer.

Nguyen, D., Hoang, S., Pham, S., & Nguyen, T. (2010). Named entity recognition for Vietnamese. *Proceedings of the Conference on Intelligent Information and Database Systems: Part II,* 205-214. Springer-Verlag.

Pasca, M., Lin, D., Bigham, J., Lifchits, A., & Jain, A. (2006). Organizing and searching the world wide web of facts – step one: the one million of fact extraction challenge. *Proceedings of the National Conference on Artificial Intelligence,* 1400-1405.

Pinnis, M. (2012). Latvian and Lithuanian named entity recognition with TildeNER. *Proceedings of the Language Resources and Evaluation,* 1258-1265.

Poibeau, T. (2003). The multilingual named entity recognition framework. *Proceedings of the Conference on European Chapter of the Associations for Computational Linguistics*, *2*, 155-158. doi: 10.3115/1067737.1067772.

Poibeau, T., & Kosseim, L. (2001). Proper name extraction from non-journalistic texts. *Proceedings of the Computational Linguistics in Netherland.* 144-157. doi: 10.1.1.21.4746.

Rajesh, S., & Goyal, V. (2011). Name entity recognition systems for Hindi using CRF approach. *Information Systems for Indian Languages: International Conference of Information Systems for Indian Languages,* 139, p. 31-35. Springer.

Ralf, S., & Bruno, P. (2007). Cross-lingual named entity recognition. *Lingvisticae Investigationes*, *30*, 135-162.

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. *Proceedings of the Conference on Computational Natural Language Learning*, 147-155.

Rushin, S., Lin, B., Gershman, A. & Frederking, R. (2010). SYNERGY: a named entity recognition system for resource-scarce languages such as Swahili using online machine translation. *Proceedings of the Workshop on African Language Technology,* 21-26.

Sari, Y., Hassan, M. F., & Zamin, N. (2009). A hybrid approach to semi-supervised named entity recognition in health, safety and environment reports. *Proceedings of the Future Computer and Communication, 599-602.*

Semenza, C. (1997). Proper-name-specific aphasias. In H.Goodglass & A.Wingfiled (Eds.), *Anomia: Neuroanatomical and cognitive correlates* (pp. 115-134). San Diego: Academic Press.

Singh, A. K. (2008). Natural language processing for less privileged languages: Where do we come from? Where are we going?. *Proceedings of the Workshop on NLP for Less Privileged Languages*, 7–12.

Singh, A. K., Pala, K., & Surana, H. (2008). Estimating the resource adaption cost from a resource rich language to a similar resource poor language. *Proceedings of the Language Resources and Evaluation*, 3514-3519.

Srihari, R.K., Niu, C. & Li, W. (2000). A hybrid approach to named entity and sub-type tagging. *Proceedings of Applied Natural Language Processing,* 247-254.

Stern, R., & Benoit, S. (2010). Resources for named entity recognition and resolution in news wires. *Proceedings of the Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management.*

Sujan, S., Sarkar, S., & Mitra, P. (2008). A hybrid feature set based maximum entropy Hindi named entity recognition. *Proceedings of the International Joint Conference in Natural Langauge Processing,* 343-350.

Szarvas, G., Farkas, R., & Kocsor, A. (2006). A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. *Discovery Science,* 267-278. Springer Berlin/Heidelberg.

Tanenblatt, M., Coden, A., & Sominsky, I. (2010). The Concept Mapper approach to named entity recognition. *Proceedings of the Language Resources and Evaluation,* 546-551.

Thelen, M., & Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 10*, 241-221.

Tongtep, N., & Theeramunkong, T. (2011). Simultaneous character-cluster-based word segmentation and named entity recognition in Thai language. *Knowledge, Information, and Creativity Support Systems: Proceedings of Knowledge, Information and Creativity Support Systems*, 6746, 216-225. Springer.

Tran, Q. T., Pham, T. T., Ngo, Q. H., Dinh, D., & Collier, N. (2007). Named entity recognition in Vietnamese documents. *Progress in Informatics Journal*, *5*, 14-17.

Utsuro, T., & Sassano, M. (2000). Minimally supervised Japanese named entity recognition: resources and evaluation. *Proceedings of the Language Resources and Evaluation*, 1229-1236.

Wan, X., Zong, L., Huang, X., Ma, T., Jia, H., Wu, Y., & Xiao, J. (2011). Named entity recognition in Chinese news comments on the web. *Proceedings of the International Joints Conference of Natural Language Processing,* 856-864.

Whitelaw, C., & Patrick, J. (2003). Evaluating corpora for named entity recognition using character-level features. *Proceedings of the Advances in Artificial Intelligence,* 910-921.

Wu, D., Sun Lee, W., Ye, N., & Leong Chieu, H. (2009). Domain adaptive bootstrapping for named entity recognition. *Proceedings of the Empirical Methods in Natural Language Processing,* 1523-1532.

Wu, Y., Zhao, J., & Xu, B. (2003). Chinese named entity recognition combining a statistical model with human knowledge. *Proceedings of the Multilingual and Mixed-language Named Entity Recognition,* 65-72.

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.

Zamin, N., & Oxley, A. (2011). Building a corpus-derived gazetteer for named entity recognition. *Proceedings of the Communications in Computer and Information Science*, 73-80.

Zamin, N., Oxley, A., Bakar, Z. A., & Farhan, S.A. (2012a). A statistical dictionary-based word alignment algorithm: An unsupervised approach. *Proceedings of the International Conference of Computer and Information Sciences,* 396-402.

Zamin, N., Oxley, A., Abu Bakar, Z., & Farhan, S. (2012b). A lazy man's way to part-of-speech tagging. *Proceedings of Knowledge Management and Acquisition for Intelligent Systems*, 106-117.

Zhou, G., & Su, J. (2002). Named entity recognition using an HMM-based chunk tagger. *Proceedings of the Annual Meeting on Association for Computational Linguistics*, 473-480.