5

Towards Domain Independent Named Entity Recognition

Fredrick Edward Kitoogo, Venansius Baryamureeba and Guy De Pauw

Named entity recognition is a preprocessing tool to many natural language processing tasks, such as text summarization, speech translation, and document categorization. Many systems for named entity recognition have been developed over the past years with substantial success save for the problem of being domain specific and making it difficult to use the different systems across domains. This work attempts to surmount the problem by proposing the use of domain independent features with a maximum entropy model and a multiobjective genetic algorithm (MOGA) to select the best features. The methods used in this work are backed up by experiments of which the classifications are evaluated using two diverse domains. Conclusions are finally drawn and the outlook for future work is considered.

1. Introduction

The rising need of automatic recognition and classification of information from different literature sources has given rise to the development of many named entity recognition systems. The systems that have emerged over the years have however been tailor-made to work with specific domains such as the newswire in the CoNLL 2003 shared task [Tjong Kim Sang and De Meulder 2003]. The systems are designed with respect to unique semantics, domain specificity, document genre and syntax which render it very difficult for adaptation across domains.

The focus of this work is placed on recognition and classification of text from diverse domains. We propose the following approach to handle the task: first, we identify the different diverse domain corpora from which the recognition is going to be done , second, we make out the entity types that are common to the different domains; in this work we adopted the entity types used in the CoNLL 2003 shared task [Tjong Kim Sang and De Meulder 2003] and those used in the MUC-7 Named Entity Task, third, we introduce the idea of domain independent features (features derived from different domain data sets) on top of local and global features. We use a maximum entropy model together with a MOGA to choose the best features performance on all domain data sets. Performance improvement with the use of a domain independent gazetteer automatically generated using the method used in Nadeau et al. [2006] will also be investigated.

The recognition and classification of entities in the approach above will be achieved in a two step process; (i) Named Entity Detection [NED], where the named entities (NEs) in the target text are tagged using the illustrious BIO model used by Tjong Kim Sang [2002] and originally proposed by Ramshaw and Marcus [1995], where a tag shows that a word is at the beginning of a NE (B), inside a NE (I) or outside a NE (O) and (ii) Named Entity

Classification [NEC] where the previously discovered named entities are then classified into predefined class types such as person name, location name, and organization name.

The novelty in our approach is that domain independence will be engineered by not only utilizing local context of a word in a sentence and other occurrences of the word in the same document, but also makes use of occurrences of each word within other domain documents to extract features (domain independent features) and the use of a multiobjective algorithm to choose out the most optimal domain independent features.

Evaluation of the approach well be realized by comparing the classification performance without using the domain independence techniques with that when the domain independence techniques have been applied over the various domain data sets. The improvement in performance using the latter demonstrates the feasibility of the approach.

1.1. Organization

The organization of this paper is as follows: after this introduction, a brief of related work is given in Section 2; The Domain Independent Named Entity Recognition System (DINERS) which includes the Data, Algorithms and Features used is illustrated in Section 3. In Section 4 the different experiments and results obtained are examined and discussed. Finally we conclude in Section 5 with a summary of the most important observations and an outlook on future work.

1.2. Related Work

The goal of the NER aims at automatically and robustly annotating named entities in large volumes of text. NER systems are required to offer good performance by being able to adapt to different domains and document genre's without much (or any) tuning. Many attempts have been made by existing NER researchers to develop systems that can successfully be tuned to new domains and applications using both hand-crafted and semi-automatic methods, however, there have been few successes in developing systems that are robust enough to automatically adapt across domains. The adaptability is mainly mired by lack of ontology and rule bottlenecks [Bontcheva et al. 2002].

Our work builds on the state-of-the-art approaches by other researchers such as Giouli et al. [2006], where in achieving domain independence their main focus was on building a homogenous, reusable and adaptable linguistic resource to different domains and languages.

On the other hand Jiang and Zhai [2006] present several strategies for exploiting the domain structure in training data to learn a more robust named entity recognizer that can perform well on a new domain. They improvise a way to automatically rank features based on their how generalizable they are across domains. They then train a classifier with strong emphasis on the most generalizable features.

Nadeau et al. [2006] use an un-supervised strategy for domain independence by creating a system that can recognize named-entities in a given document without prior training by using automatically generated gazetteers and later resolving ambiguity.

In their work, Bontcheva et al. [2002] presented an approach to domain independent named entity recognition which is portable and was built to be usable in many different applications, on many different kinds of text and for many different purposes. They specifically showed how the system is applied for annotating (semi)-automatically digital library content and also for indexing such content by entities and events.

Most of the approaches explored employed the use of local content to design features; our approach extended to exploit external content of words in different domain text and used a strategy of MOGA that is used to empirically make a choice of the most useful features.

In this work we have used some methods which are orthogonal to those used in the related work, implying that combination of the methods is capable of improving performance of domain independence in NER.

2. The Domain Independent Named Entity Recognition System (Diners)

The Domain Independent Named Entity Recognition System (DINERS), an innovation by the authors, was developed with the capability of a machine learning approach to named entity recognition to be ported from one domain to another with very limited re-engineering. The steps followed to achieve the domain independence functionality as shown in Figure 1 below and the following detail.

1.1. The Data

In this work we use English texts from the following domains:-

2.0.1. CoNLL 2003

The English set of the CoNLL 2003 data was used for this work; the CoNLL 2003 data was originally taken from the Reuters Corpus. This corpus consists of Reuters news stories. This data set represents the general domain of news stories.

The data used was already subjected to some linguistic preprocessing and files were already formatted into four columns separated by a single space. Each word has been put on a separate line and there is an empty line after each sentence. The first item on each line is a word, the second a part-of-speech (POS) tag, the third a syntactic chunk tag and the fourth the named entity tag. The chunk tags and the named entity tags have the format I-TYPE which means that the word is inside a phrase of type TYPE. Only if two phrases of the same type immediately follow each other, the first word of the second phrase will have tag B-TYPE to show that it starts a new phrase. A word with tag O is not part of a phrase. An illustration follows in Table I.

2.0.2. Courts of Judicature (Uganda)

This data was extracted from court case judgments from the Uganda Courts of Judicature mainly divided into three categories; Supreme Court, Court of Appeal and High Court. Two files of each type were used and divided into three zones, namely the Title Line (TL), Litigant Line (LL), Presiding Judicial Officer(s) Line (PJL) and Date Line (DL).

2.1. Data Processing

The different data sets were received in different formats; some were already tagged for Part of Speech (PoS), chunk tags and named entity tags, and others were not, those not tagged were manually tagged for all the three categories by annotators from AfLaT [2008].

The different data sets were further prepared by extracting useful features from which the actual learning by the machine learning algorithm is done; the final training and testing files are built in the format shown in Table II.

2.2. Entity Types to recognize

The NE guidelines developed for the CoNLL 2003 shared task [Tjong Kim Sang and De Meulder 2003] are used as the basis for the decision of the classes of NEs used in this work. This work is however limited to Person Names, Organization Names and Location Names

Fig. 1: The Domain Independent Named Entity Recognition System(DINERS) Flowchart

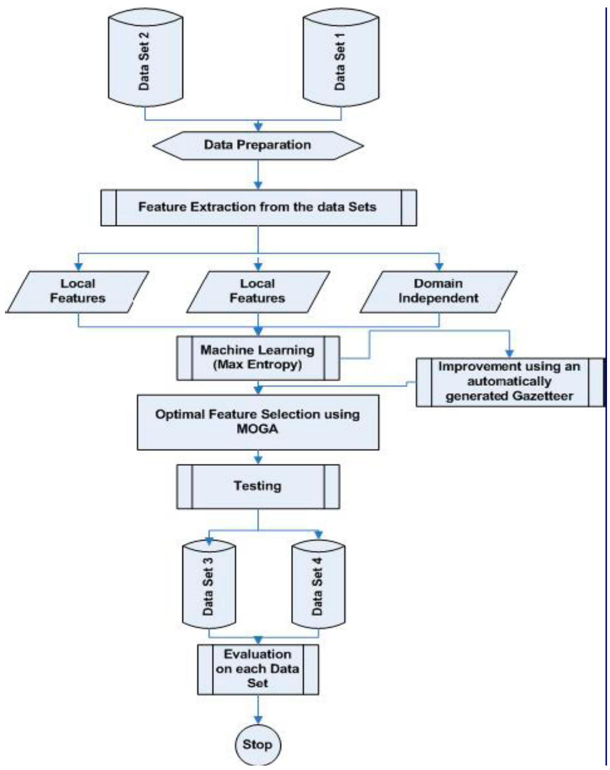

Table II. Format of the final Testing and Training files

Name Classes along with the Features				
Class	**Feature1**	**Feature2**	....	**Feature N**
Class1	binaryValue11	binaryValue21	....	binaryValue2N
..	..	..	....	....
..	..	..	....	....
Classn	binaryValue1n	binaryValue2n	....	binaryValuenN

Table I. Format of the CoNLL 2003 Data Set

Words along with the corresponding tags			
Word	**PoS Tag**	**Chunk Tag**	**Feature N**
U.N.	NNP	I-NP	I-ORG
official	NN	-NP	O
Ekeus	NNP	-NP	I-PER
heads	VBZ	-VP	O
for	IN	-PP	O
Baghdad	NNP	-NP	I-LOC
.	.		O

2.3. Feature Derivation

Features are different types of knowledge sources which are used to make tagging decisions; they are implemented a binary valued functions which query the history information in the training files and the future in the testing files to determine whether or not they fire.

2.3.1. Useful Lists

Useful lists will be generated some from without the training data while others will be derived from the training data, the following lists are derived using automated methods:-

- Frequent Word (FW) - Words that appear in more than five documents
- Rare Words (RW) - Words that appear less than five times in a document
- Useful Unigrams for each name class [UNI] - For each of the Named Entity classes (single words that come before a name); such as Honorable, Mr., Justice etc.
- Useful Bigrams for each name class [BIG] - For each of the Named Entity classes (two words that come before a name); such as His Worship, His Lordship

- Useful Name Class Suffixes (UNC) - For each of the Named Entity classes, a list of tokens that terminate them are determined; e.g. for the ORG Class we can have tokens such as Company, Limited, Inc, etc.
- Function Words for each name class [FUN] - Lower case words that occur in a name class; e.g. of the
- Gazetteer List - an important reference for information about known names

Amongst all the features some will be used for the detection stage while others will be for the classification phase. The detailed list of features is shown in Table III.

Table III. Features for Named Entity Detection and classification

Named Entity Detection Features			
Feature	**Local**	**Global**	**Domain Independent**
Rare Words - Token word found in the list of Rare Words	√		
Frequent Words - Token word found in the list of Rare Words	√		
Contextual	√		
Anchor word	√		
([-1,....+1]) Window	√		
Orthographic			
InitCap of Anchor Word	√		
Capitalization of whole Anchor word	√		
InitCap of [-1,....+1] Window	√		
Capitalization of [-1,....+1] Window	√		
Position of Anchor Word in Sentence	√		
Contains digits	√		
Contains Dollar Sign	√		
Hyphenated	√		
Proportion of capitalization of token word in the document	√	√	
Proportion of capitalization of token word in another domain documents			√
Proportion of InitCaps of token word in the document		√	
Proportion of InitCaps of token word in another domain document			√
Part of Speech Features			
PoS of token word	√		
PoS of token [-2,....+2] Window	√		
Class Suffix			
Suffix to Anchor word has UNC in the same document	√		
occurrence of the token word appears with the same UNC in the same document		√	
occurrence of the token word appears with the same UNC in another domain document			√
Unigrams			
Anchor Word has a unigram appearing in the UNI list	√		
occurrence of the token word appears with the same UNI in the same document		√	
occurrence of the token word appears with the same UNI in another domain document			√
Bigrams			
Anchor Word has a bigram appearing in the BIG list	√		
occurrence of the token word appears with the same BIG in the same document		√	
occurrence of the token word appears with the same BIG in another domain document			√
Function Word - Anchor words have FUN in inside it	√		
Lexicon Features			
Lexicon feature of Anchor Word	√		
Lexicon feature of [-1,..,+1] Window Word	√		
Anchor word is in the Gazetteer List	√		
Anchor word is in the Trigger List	√		
Anchor word is in the Name List	√		

2.1. The Classification Algorithm

The algorithm used in this work is a Maximum Entropy (ME) model based on the Kullback-Leibler divergence described by Le [2004]. The model uses the form shown in Equation 1 below.

$$p(y|x) = \frac{1}{Z(x)} exp\left[\sum_{i=1}^{k} \lambda_i f_i(x,y)\right] \quad (1)$$

Where; $p(y|x)$ denotes the conditional probability of predicting an outcome y given the context x, $f_i(x,y)'$ are the *features* with the associated weighting parameters $\lambda_i's$. k is the number of features and $Z(x)$ is a normalization factor to ensure that $\sum_y p(y|x) = 1$.

The ME model represents evidence with contextual predicates in the form in Equation 2.

$$f_{cp,y'}(x,y) = \begin{cases} 1, & \text{if } y = y' \text{ and } cp(x) = true \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Where cp is the *contextual predicate* that maps a pair of *outcomey* and *contextx* to $\{true, false\}$.

A typical example is given in Equation 3 below.

$$f_j(x,y) = \begin{cases} 1, & \text{if } word(y) = Fred \text{ and } y = \text{I - PER (or B - PER)} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

$word(x) = Fred$ is an example of a *contextual predicate*. Where I - PER means the text lies within a person name or B - PER meaning the text begins a person name.

The choice of the estimation and weighting method (Limited-Memory Variable Metric (L-BFGS))was made using the backing of Malouf [2002], in which it was empirically discovered that the Limited-Memory Variable Metric (L-BFGS) algorithm of Benson and More' [2001] outperformed the generalized Iterative Scaling, improved Iterative Scaling function optimization algorithms such as conjugate gradient and variable metric methods .

The maximum entropy classifier is used to classify each word as one of the following: the beginning of a NE (B tag), a word inside a NE (I tag), and a word outside the tag (O tag).

During experiments, it is possible that the classifier produces a sequence of inadmissible classes (e.g., PER-B followed by LOC-L). To eliminate such sequences, we adopt from Chieu and Ng [2003] the use of a transition probability between word classes P(ci—cj) to be equal to 1 if the sequence is admissible, and 0 otherwise. The probability of the classes $c_1, ..., c_n$ assigned to the words in a sequence s in a document D given other documents OD_js is defined in Equation 4 that follows:-

$$p(c_1, ...c_n|s, D, OD_j) = \prod_{i=1}^{n} P(c_i|s, D) * P(c_i|c_{i-1}) \quad (4)$$

Where $P(c_i|s, D, OD_j)$ is determined by the maximum entropy classifier. A Viterbi algorithm is then used to select the sequence of word classes with the highest probability.

2.2. The Multiobjective Genetic Algorithm (MOGA)

The multiobjective algorithm will ideally be used to select out of a pool of all the designed domain independent features an optimal set. In this work we adopted the MOGA that was used in the work of Kitoogo and Barya [2007].

2.3. Named Entity Recognition

2.3.1. Named Entity Detection

This stage is for identifying a word or sequences of words that make up the names of specific entities using the capitalization feature and a tagger to spot proper nouns. Disambiguation will be performed to solve the problem of mistaken capitalized words for named entities using the method applied in the work of Nadeau et al. [2006]. The major disambiguation tasks at this stage will be; Entity - Noun and Entity - Boundary.

Table IV. Data Sets used in the Experiments

Data Sets along with the Instances	
Data Set	**Instances**
Uganda Judiciary	
High Court 1	5,006
High Court 2	5,457
Court of Appeal 1	3,070
Court of Appeal 2	4,451
Supreme Court 1	3,156
TOTAL	21,140
CoNLL 2003	
English (Train Set)	51,577
English (Test Set A)	46,665
English (Test Set B)	204,566
TOTAL	302,808

Table V. Gazetteers

Name lists along with the Instances	
Data Set	**Instances**
Magistrates	135
Registrars	17
Judges	42

2.3.2. Named Entity Classification

At this stage, the already recognized entities are categorized into predefined classes a maximum entropy classifier using the identified feature then thereafter perform disambiguation (Entity - Entity) as in Nadeau et al. [2006].

3. Experiments

The experiments conducted in this work have been guided by the following framework and are by no means fully exhaustive; they are ideally for exploring the feasibility of the domain independence concept and DINERS specifically.

3.1. Implementation

3.1.1. Toolkit

As an implementation basis, we have used MaxEnt (Maximum Entropy Modeling Toolkit for Python and C++) toolkit developed by Le [2004], a software toolkit designed to ease the development of systems working on maximum entropy modeling [Wang et al. 2006; Hendrickx and Bosch 2004; Hendrickx and Bosch 2005]. MaxEnt offers routines for conditional maximum entropy modeling, parameter estimation and smoothing amongst others.

3.1.2. Data

We used five real-world data sets from the judgments from the Uganda Courts of Judicature and three real world data sets from the CoNLL 2003 shared task [Tjong Kim Sang and De Meulder 2003]. The two data groups (Judicial data sets and the CoNLL data sets) were integrated to form two grand data sets which were used for both training and testing interchangeably. Details of the data sets are given in Table IV.

3.1.3. Gazetteer

A list of Magistrates, Registrars and that of Judges (Justices) of the Uganda Courts of Judicature were used as the gazetteers in these experiments. A summary of the name lists is given in Table V.

Table VI. Baseline Results

Results - CoNLL 2003 used as the Training Data Set			
NAMED ENTITY	**PRECISION** (%)	**RECALL** (%)	$\mathbf{F}_{\beta=1}$ (%)
PERSON	91.35	91.87	91.61
LOCATION	90.23	90.05	90.14
ORGANIZATION	81.55	80.20	80.87
OVERALL	**90.44**	**91.18**	**90.81**

Results - Judicial Text used as the Training Data Set			
NAMED ENTITY	**PRECISION** (%)	**RECALL** (%)	$\mathbf{F}_{\beta=1}$ (%)
PERSON	70.12	71.30	70.71
LOCATION	68.93	68.21	68.57
ORGANIZATION	66.04	65.84	65.94
OVERALL	**69.55**	**69.16**	**69.35**

3.2. Experimental Methodology

The two domain data sets (CoNLL 2003 and Judicial Text) are each used as the training set as well as the testing set interchangeably; i.e first the CoNNL 2003 is used as the training set and the model is tested on the Judicial text then vice-versa. The model is trained with 100 iterations of the L-BFGS method.

The first experiments do not have domain independent features employed, while in the second experiments the domain independent features have been used together with a MOGA to test for performance improvement. The experiments are further run to test the effect of the usage of gazetteers for the PERSON entity.

The performance of the DINERS is evaluated using standard precision (P), recall (R), and F-score, where F-score is defined as 2PR/(P + R).

For comparison of performance between the use of the different feature options ($p < 0.01$), we employed the McNemar's significance tests [Dietterich 1998].

3.3. Results

The baseline results without domain independent features are shown in Table VI. Table VII in turn shows the results when domain independent features have been used. Clearly the use of domain independent features has a positive significant impact on both precision and recall rates across all the entities in both cases, i.e. even if the training and test sets are interchanged.

The experiments reveal that there is a drop in the overall f-score performance for 90.81% to 69.35% and 92.04% to 70.27% respectively for both the baseline results and domain independent features case when the data sets are swapped for training and testing. Another clear finding is that the CoNLL 2003 data set yields significantly better results when used as a training set.

As shown in Table VIII, the use of gazetteers indicates an improvement from 94.15% to 94.90% and 76.31 to 77.03 respectively, the difference in both cases is not statistically significant.

Table VII. Results - Application of Domain Independent Features

Results - CoNLL 2003 used as the Training Data Set			
NAMED ENTITY	**PRECISION** (%)	**RECALL** (%)	$\mathbf{F}_{\beta=1}$ (%)
PERSON	94.08	94.22	94.15
LOCATION	91.69	91.06	91.37
ORGANIZATION	83.78	84.05	83.91
OVERALL	**91.90**	**92.19**	**92.04**
Results - Judicial Text used as the Training Data Set			
NAMED ENTITY	**PRECISION** (%)	**RECALL** (%)	$\mathbf{F}_{\beta=1}$ (%)
PERSON	75.89	76.74	76.31
LOCATION	69.65	69.33	69.49
ORGANIZATION	67.25	66.43	66.84
OVERALL	**70.26**	**70.28**	**70.27**

Table VIII. Results - Gazetteers used with Domain Independent Features

Results - CoNLL 2003 used as the Training Data Set			
NAMED ENTITY	**PRECISION** (%)	**RECALL** (%)	$\mathbf{F}_{\beta=1}$ (%)
PERSON	95.03	94.78	94.90
Results - Judicial Text used as the Training Data Set			
NAMED ENTITY	**PRECISION** (%)	**RECALL** (%)	$\mathbf{F}_{\beta=1}$ (%)
PERSON	76.07	78.01	77.03

4. Conclusions

This work began with the identification of the different corpora, from which named entities were to be recognized, it was discovered that technical domains already annotated with named entities such as person name, location name, and organization name are not available and this necessitated manual annotation of some of the data used for evaluation.

In this work one of the initial steps of the identifying, manual annotation, and processing of judicial text was a leap towards an attempt to building a Judicial Corpora for named entity recognition. The judicial data set together with the CoNLL 2003 data set were used for evaluating the Domain Independent Recognition System (DINERS) developed in this research.

Our DINERS demonstrates that the use and empirically crafted combination of domain independent feature sets using optimization techniques such as MOGA yields improved performance. It can further be concluded from this work that non-technical data offers better training than technical data sets. The use of gazetteers for domain independence does not significantly improve performance.

Although the results reported in this work are affirmative, many research directions in this arena are yet to be explored. For future work, the system may be experimented for portability with other technical domains; the range of the domain independent features could also be widened. The use document sectioning (Title, Date, Body, etc.) will also be explored for possible performance improvement.

This work targeted generic named entities of Person, Location and Organization because it was difficult to identify technical entities that are shared between different technical domains, however for future work the work will be extended to a wider spectrum of technical domains to identify and work towards shared technical entities.

Acknowledgments

The authors wish to thank AfLaT [2008] for providing some of the data that was used for the experiments and Naomi Maajabu at AfLaT [2008] for the manual annotation of some of the judicial text data that have supported this work.

References

AFLAT. 2008. African Language Technology (AfLat). http://aflat.org/

BENSON, J.S. AND MORE, J.J. 2001. A limited memory variable metric method for bound constrained minimization. Preprint

ANL/ACSP909-0901, Argonne National Laboratory.

BONTCHEVA, K., MAYNARD, D., CUNNINGHAM, H. AND SAGGION, H. 2002. Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, 613 – 625.

CHIEU, H. L. AND NG, H. T. 2003. Named entity recognition with a maximum entropy approach. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* 4, 160 – 163.

CHINCHOR, N. 1998. MUC-7 Named Entity Task Definition (Version 3.5). In *Proceedings of the Seventh Message Understanding Conference*.

DIETTERICH, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 10(7), 1895 – 1923.

ETZIONI, O., CAFARELLA, M., DOWNEY, D., POPESCU, A.M., SHAKED, T., SODERLAND, S., WELD, D. S., AND YATES, A. 2005 Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence,* 165, 91 -134.

GIOULI, V., KONSTANDINIDIS, A., DESYPRI, E., AND PAPAGEORGIOU, X. 2006. Multi-domain, Multi-lingual Named Entity Recognition: revisiting and grounding the resources issue. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC) 2006*, Genoa, Italy.

HENDRICKX, I. AND VAN DEN BOSCH, A. 2004. Maximum-entropy parameter estimation for the k-nn modified value-difference kernel. In *Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence*, 19 – 26.

HENDRICKX, I. AND VAN DEN BOSCH, A. 2005. Algorithms with Instance-based Classification. In *Proceedings of the 16th European Conference on Machine Learning*, Porto, Portugal, October 3-7 2005, 158 – 169.

JIANG, J. AND ZHAI C. X. 2006. Exploiting Domain Structure for Named Entity Recognition. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 74 – 81.

KITOOGO, F.E., AND BARYAMUREEBA, V. 2007. A Methodology for Feature Selection in Named Entity Recognition, *International Journal of Computing and ICT Research in Computing and ICT Research (IJCIR)*, 1(1), 18 – 26.

LE, Z. (2004). Maximum Entropy Modeling Toolkit for Python and C++. *Natural Language Processing Lab*, Northeastern University, China.

MALOUF, R. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the International Conference on Computational Linguistics*, 1 – 7.

NADEAU, D., TURNEY, P., AND MATWIN, S. 2006. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. In *Proceedings of 19th Canadian Conference on Artificial Intelligence*.

RAMSHAW, L. AND MARCUS, M. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, 82 – 94.

TJONG KIM SANG, E.F. 2002. Introduction to the conll-2002 shared task: Language independent named entity recognition. In *Proceedings of CoNLL-2002*, 155 – 158.

TJONG KIM SANG, E.F. AND DE MEULDER, F. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, 142 – 147.

WANG, M., SAGAE, K., AND MITAMURA, T. 2006. A fast, accurate deterministic parser for Chinese. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, 425 – 432.