# Baseline Experiments for OCR-Based Arabic Named-Entity Extraction

Tapas Kanungo and Osama Bulbul

Laboratory for Language and Media Processing
Center for Automation Research
University of Maryland at College Park
College Park, MD 20742, USA
kanungo@cfar.umd.edu

**Abstract.** The named-entity extraction task is concerned with extracting higher-level information chunks such as names of people, organization, time, and dates from messages. Most recognition systems that perform this task assume that the input to the system is noise-free symbolic text. In this paper we explore the impact of OCR error on the accuracy of named-entity recognition systems. We report results of our baseline experiments that we conducted using an off-the-shelf named-entity recognizer and a commercial OCR product. We hope that the dataset and the experimental results will stimulate cross-disciplinary research and allow researchers to measure progress each year.

## 1 Introduction

Optical Character Recognition (OCR) converts document images into editable and searchable symbolic text [3]. While OCR allows users to search for specific strings in the text, it would benefit users if they could search the symbolic text at a meta-level instead of the standard string search. For example, a user might want to know the names of all the persons mentioned in a specific document without having to search one-by-one for strings corresponding to names in a database. Similarly another user might want to know if any organization is mentioned in a document. While many named-entity tagging systems have been developed for error-free symbolic text, we explore the impact of OCR errors on these tagging systems. We hope that these baseline experiments will allow researchers to track the progress each year and stimulate cross-disciplinary research among researchers in OCR, computational linguistics, and information retrieval research areas.

The named-entity task was part of the DARPA sponsored Message Understanding program [7] and is described in their conference proceedings [4]. MITRE corporation built an environment called Alembic [1, 6] that allows users to automatically tag named-entities and evaluate the results. The SRA corporation developed a named-entity tagger for Arabic language that is based on morphological analysis of the words [13]. An evaluation/scoring system based on the task definitions and the metrics was developed at SAIC [8]. Researchers have also built named-entity extraction systems for output generated from speech recognition systems [12, 9]. Earlier [11] we reported results of our evaluation of two Arabic OCR products: Sakhr's Automatic Page Reader and OnsetTechnology's OmniPage for Arabic. While there is literature on impact of OCR errors on information retrieval performance [5, 2], we are not aware of research that addresses the issue of extracting named-entities from noisy OCR text.

In Section 2 we define the named-entity recognition task. The methodology for conducting the experiments is described in Section 3. The dataset, the OCR system, and the named-entity tagging system is described in Section 4. Finally our experimental findings are reported in Section 5.

## 2 The Named-Entity Recognition Task

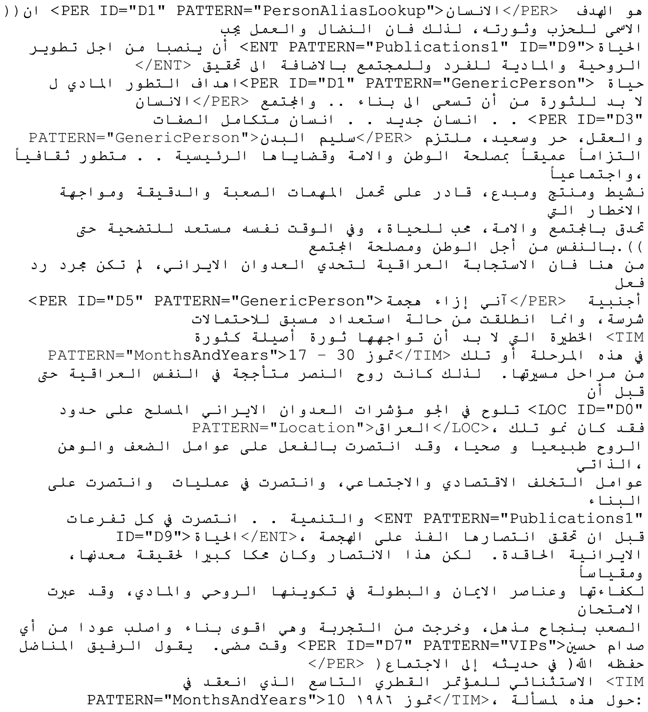
The original MUC definition of named entities [4] categorized named enties into three types: i) entity names (e.g. organizations, persons and locations), ii) temporal expressions (e.g. dates and times), and iii) numeric expressions (e.g. monetary values and percentages). However, the SRA named-entity tagger used in our experiments categorized entities into five top-level categories: numeric (monetary values, percentages), entities (organizations, publications), temporal (time, date), location (city, country, location), and person. The task was not to recognize these entities by pattern matching but by using context and linguistic structure [10]. A sample Arabic document with its SGML named-entity tags is shown in Figure 1.

## 3 Methodology

The methodology used in our experiments is illustrated in Figure 2. Please refer to this figure for rest of this section. There are two components of the entire end-to-end OCR-based named-entity recognition system. The individual components are evaluated independently and then as a combined system. Four kinds of experiments are conducted: i) Manual transcription (OCR Groundtruth) is used to evaluate the OCR character and word accuracy (OCREval Result), ii) manual named entity annotations (ManualNE) are used to evaluate automatic named-entity tagging performance (AutoNE Result), iii) automatic named-entity tags on the OCR output are compared against automatic named-entity tags on the OCR Groundtruth to isolate the tagging error introduced only due to OCR, i.e., it is assumed that the automatic tagger produced perfect tagging results, and finally iv) manual named-entity annotations are compared against the automatic named-entity tagging results on the OCR output.

## 4 Experimental Protocol

We developed an Arabic corpus for conducting end-to-end OCR-based named entity extraction experiments. A total of 115 images containing Arabic text were



```
هو الهدف </PER>الانسان<PER ID="D1" PATTERN="PersonAliasLookup"> ان((
الاسمى للحزب وثورته، لذلك فان النضال والعمل يجب
الحياة<ENT PATTERN="Publications1" ID="D9"> أن ينصبا من اجل تطوير
الروحية والمادية للفرد وللمجتمع بالاضافة الى تحقيق </ENT>
حياة <PER ID="D1" PATTERN="GenericPerson">اهداف التطور المادي ل
لا بد للثورة من أن تسعى الى بناء .. والمجتمع </PER>الانسان
<PER ID="D3" . . انسان جديد . . انسان متكامل الصفات
والعقل، حر وسعيد، ملتزم </PER>سليم البدن<PATTERN="GenericPerson"
التزاماً عميقاً بمصلحة الوطن والامة وقضاياها الرئيسية . . متطور ثقافياً
،واجتماعياً
نشيط ومنتج ومبدع، قادر على تحمل المهمات الصعبة والدقيقة ومواجهة
الاخطار التي
تحدق بالمجتمع والامة، محب للحياة، وفي الوقت نفسه مستعد للتضحية حتى
)).بالنفس من أجل الوطن ومصلحة المجتمع
من هنا فان الاستجابة العراقية لتحدي العدوان الايراني، لم تكن مجرد رد
فعل
أجنبية </PER>آني إزاء هجمة<PER ID="D5" PATTERN="GenericPerson">
شرسة، وانما انطلقت من حالة استعداد مسبق للاحتمالات
<TIM الخطيرة التي لا بد أن تواجهها ثورة أصيلة كثورة
في هذه المرحلة أو تلك </TIM>تموز 30 - PATTERN="MonthsAndYears">17
من مراحل مسيرتها. لذلك كانت روح النصر متأججة في النفس العراقية حتى
قبل أن
<LOC ID="D0" تلوح في الجو مؤشرات العدوان الايراني المسلح على حدود
فقد كان نمو تلك ،</LOC>العراق<PATTERN="Location"
الروح طبيعيا و صحيا، وقد انتصرت بالفعل على عوامل الضعف والوهن
،الذاتي
عوامل التخلف الاقتصادي والاجتماعي، وانتصرت في عمليات  وانتصرت على
البناء
<ENT PATTERN="Publications1" والتنمية . . انتصرت في كل تفرعات
قبل ان تحقق انتصارها الفذ على الهجمة ،</ENT>الحياة<ID="D9"
الايرانية الحاقدة. لكن هذا الانتصار وكان محكا كبيرا لحقيقة معدنها،
ومقياساً
لكفاءتها وعناصر الايمان والبطولة في تكوينها الروحي والمادي، وقد عبرت
الامتحان
الصعب بنجاح مذهل، وخرجت من التجربة وهي اقوى بناء واصلب عودا من أي
صدام حسين<PER ID="D7" PATTERN="VIPs"> وقت مضى. يقول الرفيق المناضل
حفظه الله( في حديثه إلى الاجتماع( </PER>
<TIM الاستثنائي للمؤتمر القطري التاسع الذي انعقد في
حول هذه لمسألة ،</TIM>تموز ١٩٨٦ PATTERN="MonthsAndYears">10:
```

**Fig. 1.** Named-entity tags produced by the SRA Tagarab system. The three-letter labels are: NUM (numeric), ENT (entity), TIM (time), LOC (location), PER (Person). Notice that the system also produces finer distinctions such as "GenericPerson" and "VIPs" under the "PER" category.

provided to us by the Department of Defense. These images are at 300 dpi resolution, and have width of 2544 pixels and a length of 3300 pixels. Each page has one column Arabic text. The document images are from one book. The content of the book is political. The averages of some of the corpus characteristics are as follows: lines/page = 24.12; number of words/page = 241.82; characters/page = 1382.46; words/page = 241.82; words/line = 10.03 ; characters/line = 57.32; length of words = 5.12.

Manual annotation of the named entities as well as creation of the transcription was done by only one linguist due to lack of resources. The SRA named-entity graphical annotation tool [13] was used for creating manual named-entity tags. A named-entity tagging guide [4] was used for manual annotation.

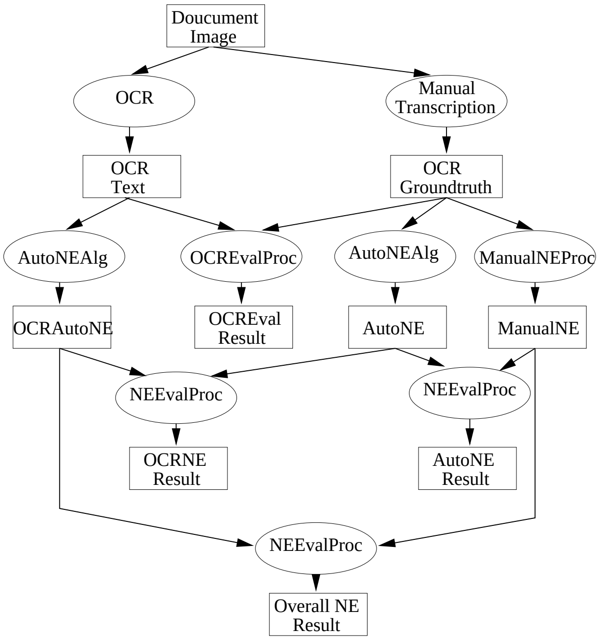


**Fig. 2.** Overall methodology used for the experiments.

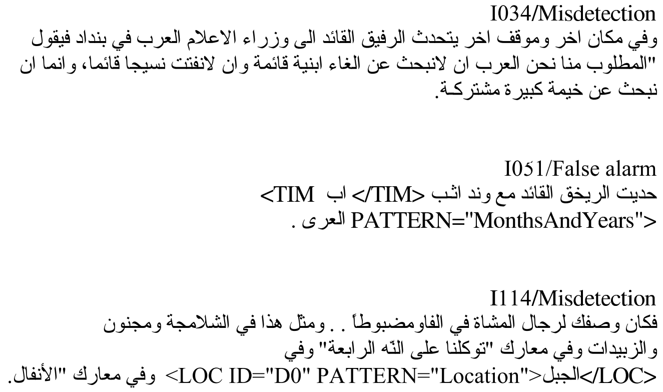
We used Sakhr's Automatic Reader version 3.01 as our Arabic OCR system. This product runs on Arabic Windows 95 and had the best character and word recognition accuracy amongst commercial Arabic OCR products [11]. The output of the OCR system is in CP1256 encoding. The definitions of the OCR evaluation metrics can be found our earlier article [11].

The Arabic named-entity recognition system we used in our experments was the Tagarab tagger developed at the SRA corporation [13]. It is based on their Turbotag English named-entity tagger and uses morphological analysis[1] of the words to tag the text. The system accepts ASMO-encoded text and marks up the named-entities text using SGML tags. The Tagarab scoring software, which is based on the MUC scoring metrics precision and recall [8], was used to evaluate the named-entity tagging performance. Partially correct recognition results are given either full, partial, or no credit and the corresponding results are reported.

## 5 Results

The OCR performance (including the 95% confidence intervals) was as follows. Average character accuracy per page: $92.3242 \pm 2.406$; average character error

[1] Here morphological analysis is in the linguistic sense and not in the image processing sense.



I034/Misdetection

وفي مكان اخر وموقف اخر يتحدث الرفيق القائد الى وزراء الاعلام العرب في بنداد فيقول
"المطلوب منا نحن العرب ان لانبحث عن الغاء ابنية قائمة وان لانفتت نسيجا قائما، وانما ان
نبحث عن خيمة كبيرة مشتركة.

I051/False alarm

حديث الريخق القائد مع وند اثب </TIM> اب <TIM
PATTERN="MonthsAndYears"> العرى .

I114/Misdetection

فكان وصفك لرجال المشاة في الفاومضبوطاً . . ومثل هذا في الشلامجة ومجنون
والزبيدات وفي معارك "توكلنا على النّه الرابعة" وفي
<LOC ID="D0" PATTERN="Location">الجبل</LOC> وفي معارك "الأنفال.

**Fig. 3.** Misdetection and false alarm examples. In the top example, 'Bagdad' is misspelt as 'Bandad" and is not detected as a 'Location' entity. In the middle example a word got split into two fragments, one of which happens to be the name of a month and so got tagged as a 'Time' entity. The word 'God' is mis-spelt in the final example and is mis-detected.

rate per page: 14.1891 ± 4.4287; average word accuracy per page is: 65.2268 ± 4.0987; and, average word error rate per page: 53.5134 ± 9.0031.

**Table 1.** Named entity recognition results of automatic tagging against manual tagging: (a) full credit, (c) half credit, and (c) no credit.

(a)

| Type | Man | Auto | Recall | Precision |
|---|---|---|---|---|
| NUM | 48 | 4 | 4.2 | 50.0 |
| ENT | 470 | 232 | 9.8 | 19.8 |
| TIM | 330 | 250 | 34.5 | 45.6 |
| LOC | 840 | 878 | 48.3 | 46.2 |
| PER | 638 | 416 | 24.1 | 37.0 |
| TOT | 2326 | 1780 | 31.0 | 40.6 |

(b)

| Type | Man | Auto | Recall | Precision |
|---|---|---|---|---|
| NUM | 48 | 4 | 3.1 | 37.5 |
| ENT | 470 | 232 | 8.7 | 17.7 |
| TIM | 330 | 250 | 30.2 | 39.8 |
| LOC | 840 | 878 | 46.0 | 44.0 |
| PER | 638 | 416 | 23.0 | 35.3 |
| TOT | 2326 | 1780 | 29.0 | 37.9 |

(c)

| Type | Man | Auto | Recall | Precision |
|---|---|---|---|---|
| NUM | 48 | 4 | 2.1 | 25.0 |
| ENT | 470 | 232 | 7.7 | 15.5 |
| TIM | 330 | 250 | 25.8 | 34.0 |
| LOC | 840 | 878 | 43.6 | 41.7 |
| PER | 638 | 416 | 21.9 | 33.7 |
| TOT | 2326 | 1780 | 27.0 | 35.3 |

In Figure 3 we show examples of named-entity tagging errors. The performance of Tagarab on clean text is shown in Table 1. In each table, the type of entity being considered is reported in the first column, the numbers in the first and second columns represent the number of tags of each type that are found by the two methods being compared, and the third and fourth column represent recall and precision, respectively. The total number of tags are reported in the row labeled 'TOT'. The performance of Tagarab on named-entity tagging of the OCR text assuming that the tagging on the manual transcription is perfect is shown in Table 2. Finally, the overall performance of Tagarab on the OCR text

**Table 2.** Recognition results of named-entity tagging of OCR output againt automatic named-entity tagging: (a) full credit, (c) half credit, and (c) no credit.

(a)

| Type | Auto | OCR | Recall | Precision |
|---|---|---|---|---|
| NUM | 2 | 0 | 0.0 | - |
| ENT | 176 | 166 | 48.9 | 51.8 |
| TIM | 266 | 204 | 40.6 | 52.9 |
| LOC | 910 | 688 | 29.5 | 39.0 |
| PER | 426 | 666 | 36.6 | 23.4 |
| TOT | 1780 | 1724 | 34.7 | 35.8 |

(b)

| Type | Auto | OCR | Recall | Precision |
|---|---|---|---|---|
| NUM | 2 | 0 | 0.0 | - |
| ENT | 176 | 166 | 47.2 | 50.0 |
| TIM | 266 | 204 | 36.1 | 47.1 |
| LOC | 910 | 688 | 28.2 | 37.4 |
| PER | 426 | 666 | 34.3 | 21.9 |
| TOT | 1780 | 1724 | 32.7 | 33.8 |

(c)

| Type | Auto | OCR | Recall | Precision |
|---|---|---|---|---|
| NUM | 2 | 0 | 0.0 | - |
| ENT | 176 | 166 | 45.5 | 48.2 |
| TIM | 266 | 204 | 31.6 | 41.2 |
| LOC | 910 | 688 | 27.0 | 35.8 |
| PER | 426 | 666 | 31.9 | 20.4 |
| TOT | 1780 | 1724 | 30.7 | 31.7 |

**Table 3.** Named entity recognition results of named-entity tagging of OCR output against manual named-entity tagging: (a) full credit, (c) half credit, and (c) no credit.

(a)

| Type | Man | OCR | Recall | Precision |
|---|---|---|---|---|
| Num | 48 | 4 | 0.0 | 0.0 |
| ENT | 470 | 190 | 8.5 | 21.1 |
| TIM | 330 | 208 | 21.8 | 34.6 |
| LOC | 840 | 672 | 21.2 | 26.5 |
| PER | 638 | 650 | 21.0 | 20.6 |
| TOT | 2326 | 1724 | 18.2 | 24.6 |

(b)

| Type | Man | OCR | Recall | Precision |
|---|---|---|---|---|
| Num | 48 | 4 | 0.0 | 0.0 |
| ENT | 470 | 190 | 7.6 | 18.7 |
| TIM | 330 | 208 | 16.8 | 26.7 |
| LOC | 840 | 672 | 19.5 | 24.4 |
| PER | 638 | 650 | 19.1 | 18.8 |
| TOT | 2326 | 1724 | 16.2 | 21.9 |

(c)

| Type | Man | OCR | Recall | Precision |
|---|---|---|---|---|
| Num | 48 | 4 | 0.0 | 0.0 |
| ENT | 470 | 190 | 6.6 | 16.3 |
| TIM | 330 | 208 | 11.8 | 18.8 |
| LOC | 840 | 672 | 17.9 | 22.3 |
| PER | 638 | 650 | 17.2 | 16.9 |
| TOT | 2326 | 1724 | 14.2 | 19.1 |

is computed by comparing the result against the manual named-entity tags. This is shown in Table 3.

## Acknowledgements

We would like to thank Gary Strong of the National Science Foundation for suggesting this problem; Doug Oard of the University of Maryland for helping us find the SRA named-entity tagger; Mark Mason, Boyan Onyshkevych and Glenn van Doren of the Department of Defense for providing access to the SRA TagArab named-entity tagger and the named-entity annotation tools; John Maloney of SRA for pointing me to relevant papers; and the Department of Defense for providing us with the scanned images.

This research was funded in part by the Department of Defense under Contract MDA 9049-6C-1250, Lockheed Martin under Contract 9802167270, the Defense Advanced Research Projects Agency under Contract N660010028910, and the National Science Foundation under Grant IIS9987944.

## References

1. J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson, and M. Vilain. MITRE: Description of the Alembic system used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*, pages 141–155, Columbia, MD, 1995.

2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.
3. H. Bunke and P.S.P. Wang, editors. *Handbook of Character Recognition and Document Image Analysis.* World Scientific Publishing Co. Pte. Ltd., 1997.
4. N. Chinchor. MUC-7 named entity task definition version 3.5. In *Proceedings of the Seventh Message Understanding Conference*, Washington, DC, 1997.
5. W. B. Croft, S. M. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 115–126, Las Vegas, NV, April 1994.
6. D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. Mixed-initiative development of language processing systems. In *Fifth Conference on Applied Natural Language Processing*, Washington, DC, March 1997.
7. Defense Adavanced Research Projects Agency. *Proceedings of the Sixth Message Understanding Conference*, Columbia, MD, November 1995.
8. A. Douthat. The message understanding conference scoring software user's manual. In *Proceedings of the Seventh Message Understanding Conference*, Washington, DC, 1997. `http://jaguar.ncsl.nist.gov/pub /ieeval0.3.tar.gz`.
9. V. Goel and W. Byrne. Task dependent loss functions in speech recognition: Application to named entity extraction. In *Proceedings of ESCA ETRW Workshop on Accessing Information in Spoken Audio*, 1999.
10. D. Jurafsky and J. H. Martin. *Speech and Language Processing.* Prentice Hall, 2000.
11. T. Kanungo, G. A. Marton, and O. Bulbul. OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products. In *Proceedings of SPIE Conference on Document Recognition*, volume 3651, pages 109–120, San Jose, CA, January 1999.
12. F. Kubala, R. Schwartz, R. Stone, and R. Weischedel. Named entity extraction from speech. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages Lansdowne, VA, February 1998.
13. J. Maloney and M. Niv. TAGARAB: A fast, accurate Arabic name recognizer using high-precision morphological analysis. In *Proceedings of COLING-ACL Workshop on Computational Approaches to Semitic Languages*, August 1998.