A guide to *IBN SINA Ext* Database

Reza FARRAHI MOGHADDAM and Mohamed CHERIET

…

Table of Contents

Summary

IBN SINA Ext database is an extension to IBN SINA database, which was published earlier as part of Active Learning Challenge 2010, and reported in DAS'10. In the extended database, more data, including the images of sub-words are available. Please see Description section for more detail.

Description

IBN SINA Ext database covers 50 folio of a manuscript. The list of folios is available in *docinfo.mat* file. The a priori information used is available in *apriori.mat*. Finally, there is a file for each folio which contains a structure called *block_info*, which host all the information, including the images and features of each sub-word. For example, *block_info* of folio 0011 (which its full image is provided as 0011.pdf) has the following format:

```
load 0011stdalnSM-blockinfo.mat
block_info =
     u0name: '0011stdalnSM'
     u0name_index: 1
     med_dir: '../../med/set10_complete/'
     number_components: 489
     CC: {1x489 cell}
```

CC subfields contain the sub-word information. This folio has 489 sub-words. For example, for CC=5, we have:

```
block_info.CC{5}
     main: [1x1 struct]
     meta: [1x1 struct]
     features: [1x1 struct]
     labels: [1x1 struct]
```

Subfield *main.compact* contains the image of sub-word (binarized), and subfield *main.compact_color* contains the color version of the image:

```
figure, imshow(block_info.CC{5}.main.compact)
```

```
figure, imagesc(block_info.CC{205}.main.compact_color)
```

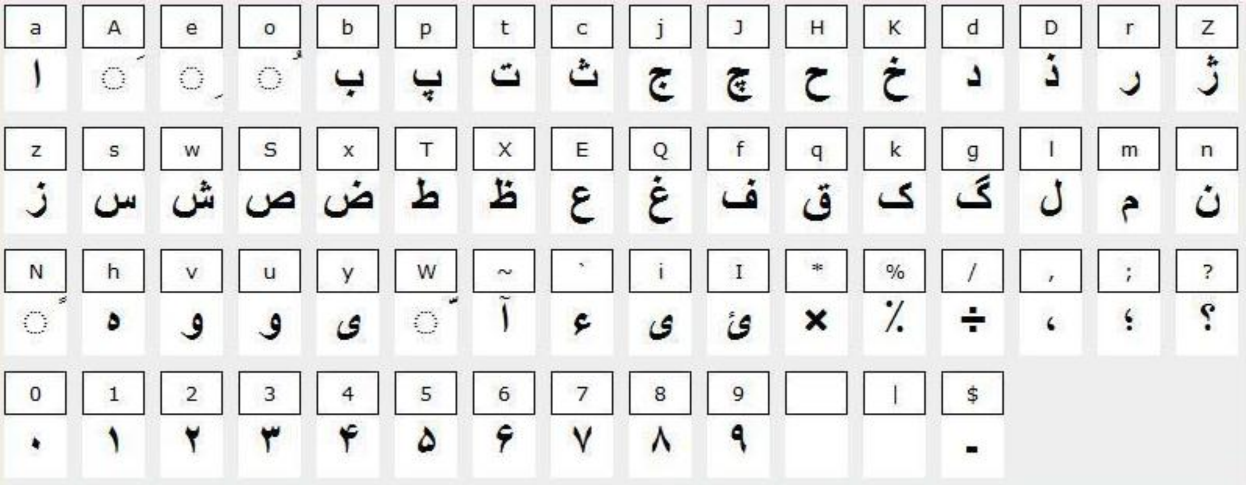
Other subfields of main are:

```
block_info.CC{5}.main
     compact: [25x56 double]
     bounds: [1277 1301 321 376]
     masscenter: [18.3048 26.8604]
     masscenter_global: [1.2953e+003 347.8604]
     dotted_feature: 0
     lib_index: {[1]  [1]}
     descriptors_raw: [1x1 struct]
     points: [1x1 struct]
     meta_features: [1x1 struct]
```

More details on various subfields of *block_info* will be added to this guide. For the moment, the other important subfield is the label of sub-word which is accessible in *labels.fstring* subfield:

```
block_info.CC{5}.labels.fstring
     sd
```

As can be seen, the labels are in Latin transliteration. We use Finglish transliteration shown in Figure 1.



a	A	e	o	b	p	t	c	j	J	H	K	d	D	r	Z
ا	◌َ	◌ِ	◌ُ	ب	پ	ت	ث	ج	چ	ح	خ	د	ذ	ر	ژ
z	s	w	S	x	T	X	E	Q	f	q	k	g	l	m	n
ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ک	گ	ل	م	ن
N	h	v	u	y	W	~	`	i	I	*	%	/	,	;	?
◌ً	ه	و	و	ی	◌ّ	آ	ء	ی	ئ	×	٪	÷	،	؛	؟
0	1	2	3	4	5	6	7	8	9		|	$			
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹			ـ			

Figure 1. Finglish table.

Acknowledgements

The authors would like to thank Prof. Robert Wisnovsky and his team at Institute of Islamic Studies, McGill University, Montreal, Canada, for their great help in manual validation of labels.

References

1. R. Farrahi Moghaddam, M. Cheriet, M. M. Adankon, K. Filonenko, and R. Wisnovsky, “IBN SINA: a database for research on processing and understanding of Arabic manuscripts images," in DAS'10. Boston, Massachusetts, 2010, pp. 11-18, DOI: 10.1145/1815330.1815332.