13th International Conference
on Document Analysis and
Recognition (ICDAR 2015)
25 August 2015

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Solving Substitution Ciphers for OCR with a Semi-supervised Hidden Markov Model

Erik Scharwächter
Stephan Vogel

Problem

- Variety of fonts, writing styles
- Commercial OCR system fails
- Training shapes for every document infeasible

Images: QNLib, ALTEC
Tirana Observer

Unsupervised OCR

- No shape training
- View OCR as cryptanalysis problem
- Dates back to 1968: R. G. Casey and G. Nagy, “An autonomous reading machine"

General procedure

1. Perform segmentation into words and characters
2. Cluster character images by shape similarity
3. Generate ciphertext from cluster identifiers
4. Solve resulting substitution cipher

Step 1: Character segmentation

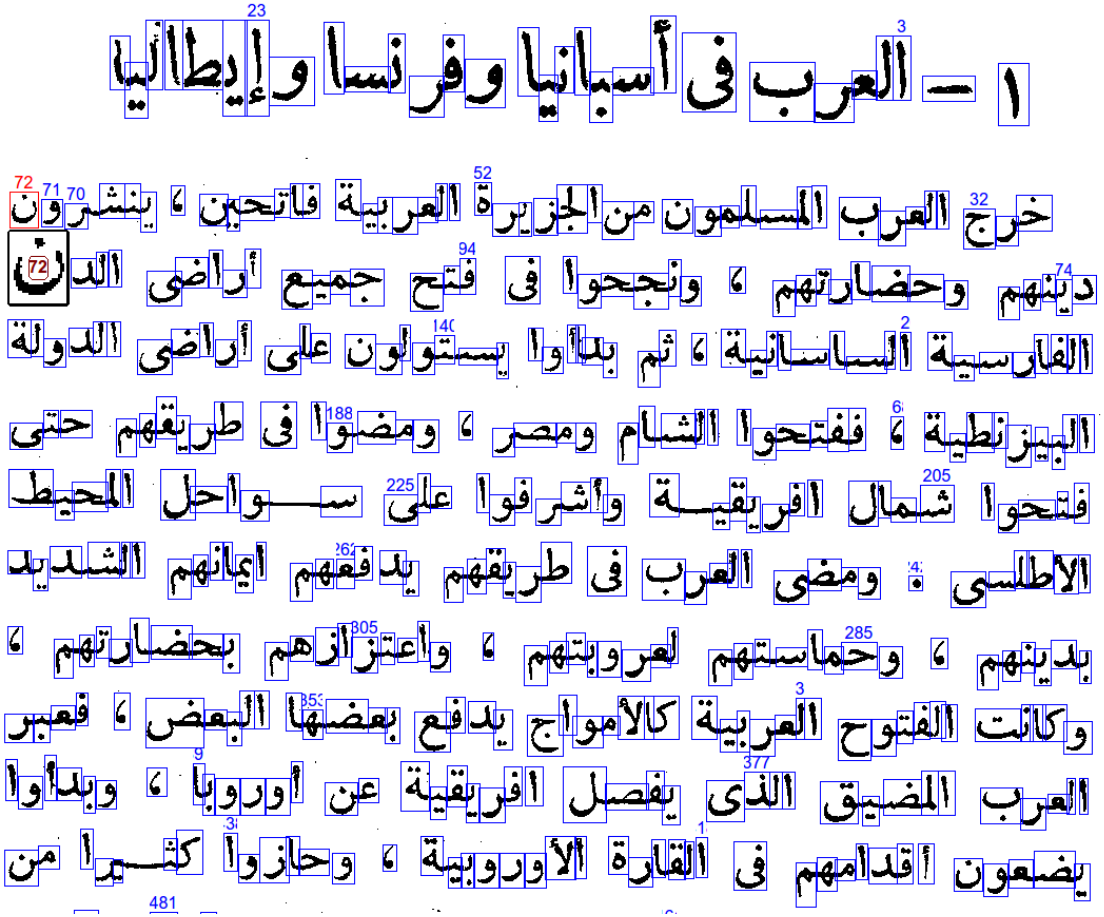
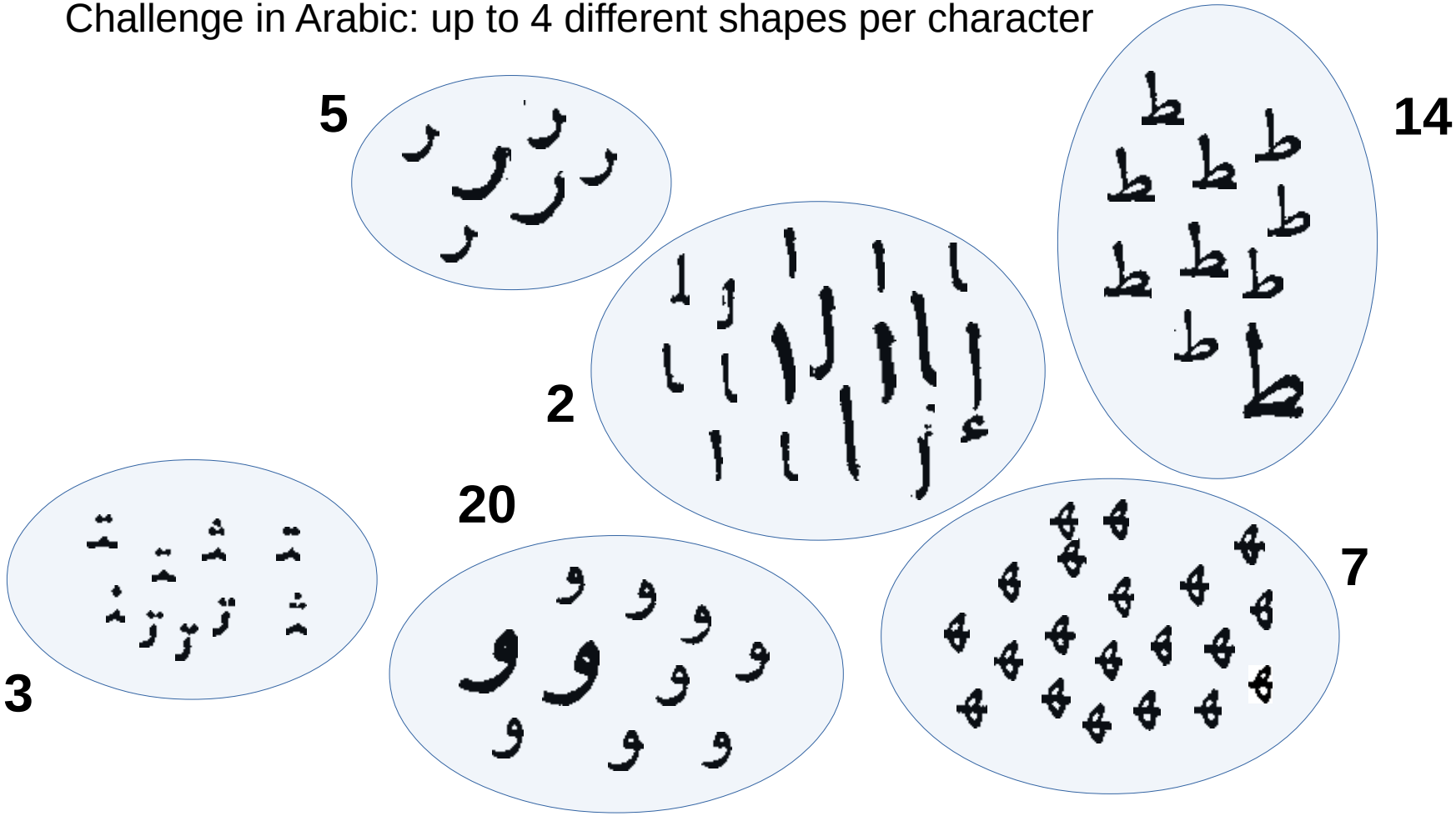


Image: ALTEC, screenshot of Cowboxer

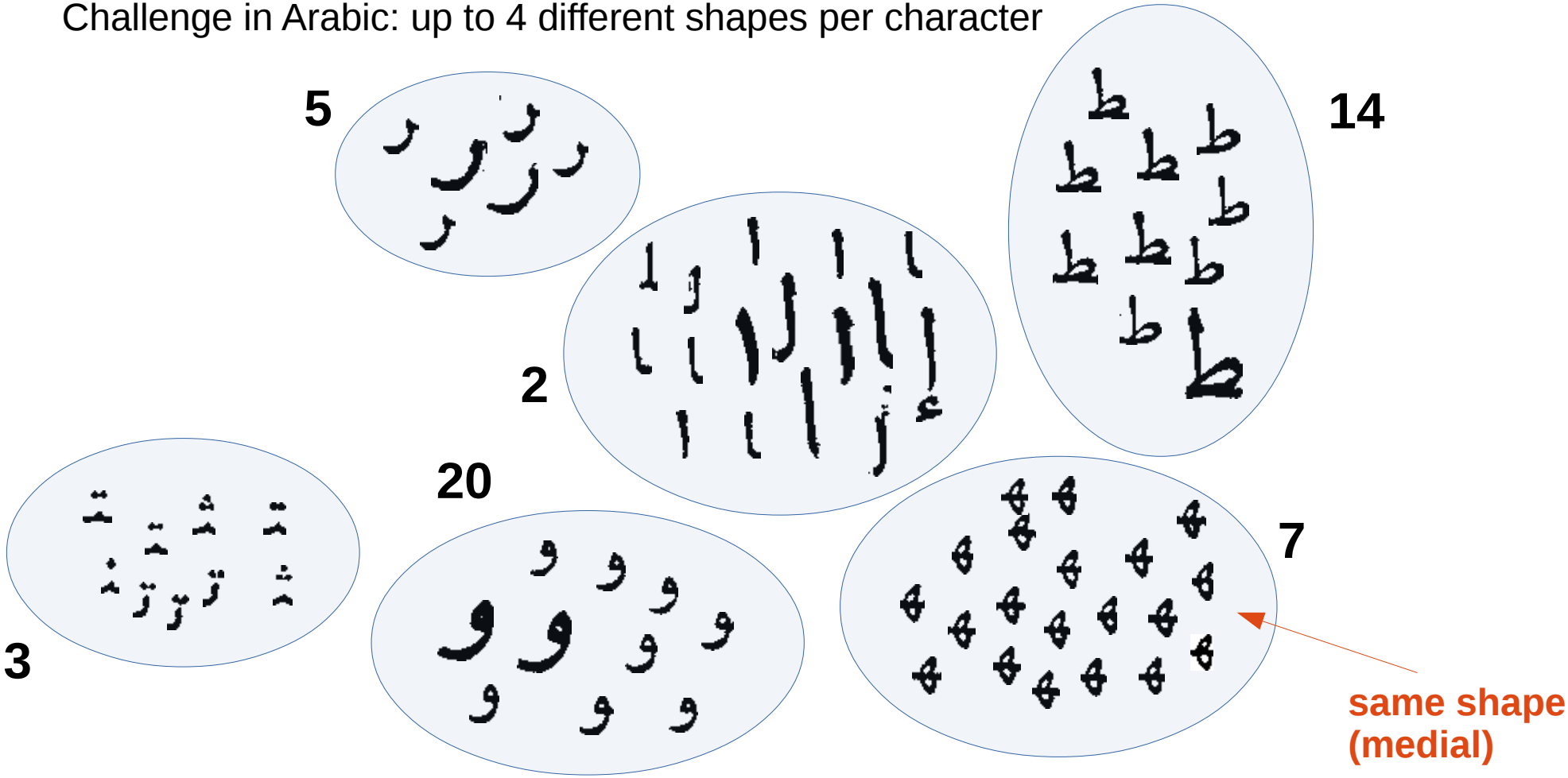
Step 2: Clustering

Challenge in Arabic: up to 4 different shapes per character



Step 2: Clustering

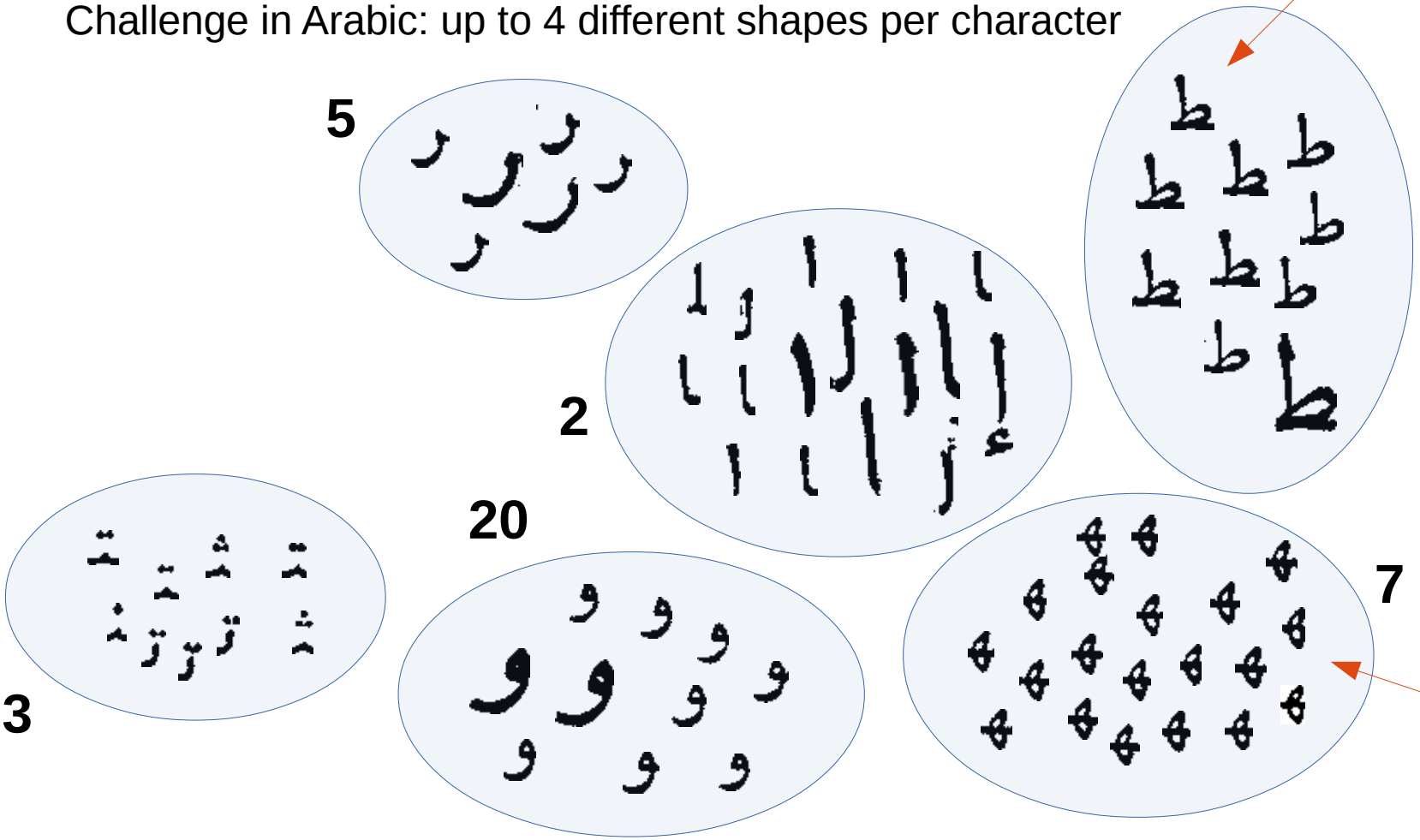
Challenge in Arabic: up to 4 different shapes per character



Step 2: Clustering

Challenge in Arabic: up to 4 different shapes per character

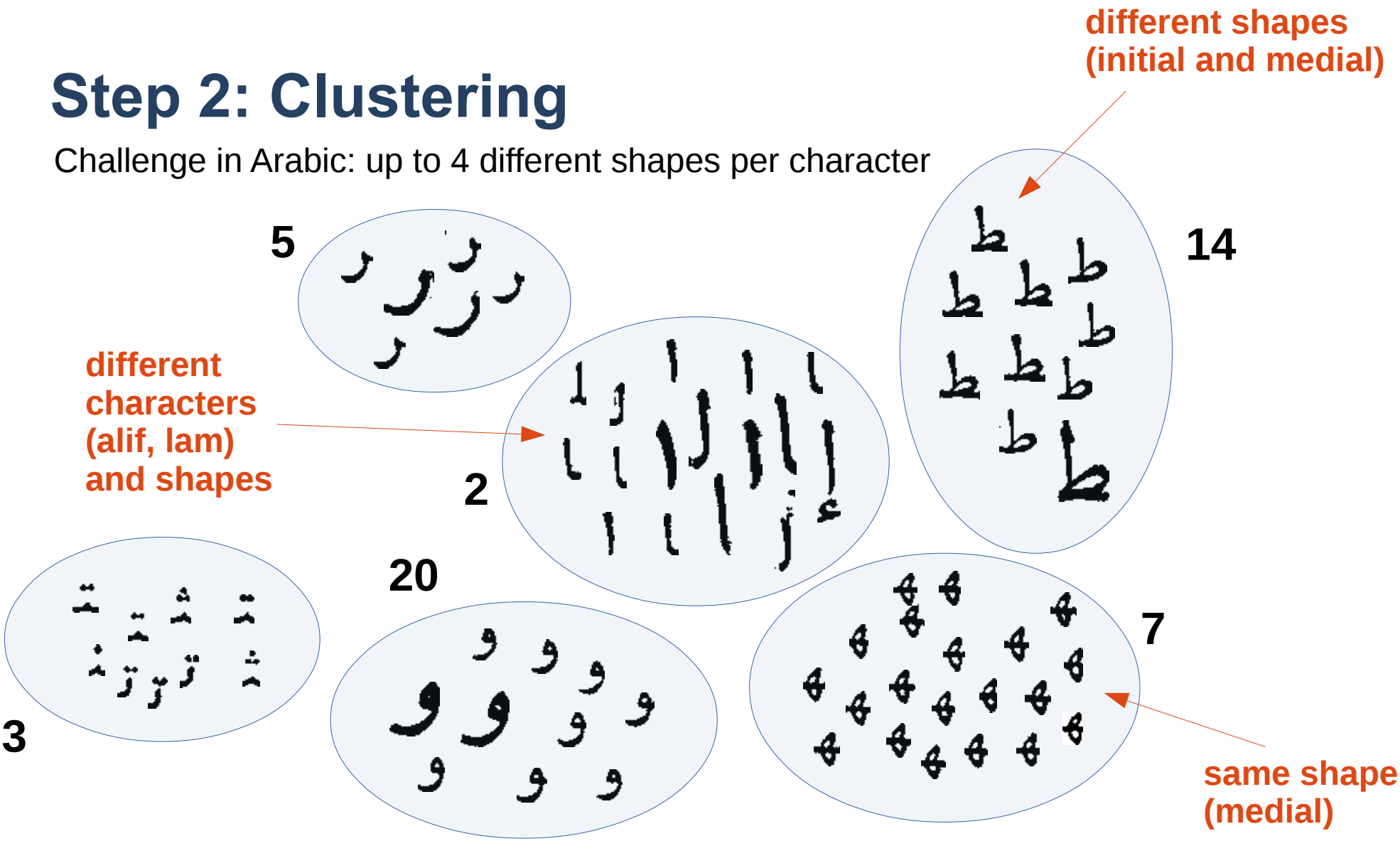
different shapes
(initial and medial)



same shape
(medial)

Step 2: Clustering

Challenge in Arabic: up to 4 different shapes per character

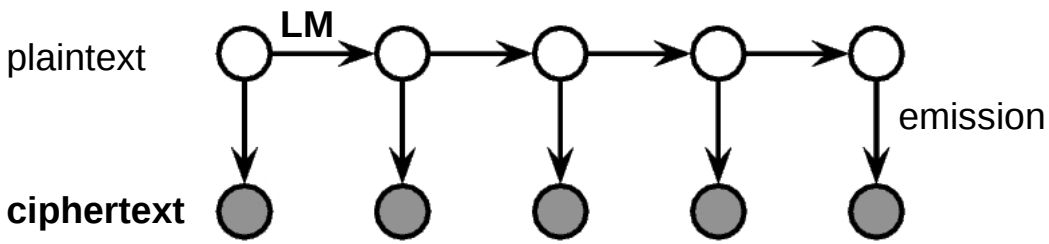


Step 3: Generate ciphertext

61 10 11 18 53 57 36 58 70 62 51 68 26 13 54 3 75 56 60 35 65 65 11 47 28 19 2 47 3 62 17 58 43 17 7 61 63 44 40
46 44 59 70 62 55 2 40 46 67 55 70 25 70 3 24 59 48 51 9 20 61 5 80 46 33 55 8 63 41 34 38 4 32 64 5 62 35 58 25 21 50 61 63 77 40
46 7 76 31 78 3 49 2 40 48 65 59 8 56 60 21 17 67 11 70 33 30 6 55 62 16 78 6 58 30 17 65 11 62 33 73 68 53 36 11 64
65 11 75 60 21 58 55 70 50 46 65 62 33 14 24 45 54 5 37 62 65 60 17 4 32 64 55 29 71 36 50 71 32 75 2 25
71 1 49 32 55 43 19 54 6 58 43 21 77 25 62 64 65 66 62 33 73 68 53 10 61 65 6 68 26 71 5 52 3 68 53 46 41 11 61 10 61
20 61 10 11 68 26 16 63 36 19 70 55 11 68 26 7 19 62 54 39 8 68 53 16 63 17 4 11 27 22 27 15 9 20 56 58 70 62 51 68 26 16
34 74 70 3 23 46 7 59 11 30 47 35 44 6 62 20 69 40 74 52 1 6 30 46 40 33 65 61 47 20 6 58 68 70 71 67 76 6 25 16 65 6 76 62
46 41 6 62 64 17 32 58 2 65 46 7 27 43 10 59 55 50 46 65 62 33 73 31 40 4 38 5 57 62 54 20 70 13 80 20 25 5 63 46
10 58 6 30 38 5 65 61 71 6 68 53 65 71 44 73 70 62 14 46 1 54 62 34 20 31 45 16 57 36 70 27 57 71 74 11 37 28 46 6 38
5 65 49 22 62 15 70 16 54 3 75 56 30 35 65 27 11 76 24 23 5 65 61 71 6 68 26 65 31 68 70 65 11 62 14 66 30 33 25 40 6 38
35 7 27 19 38 13 3 8 58 80 71 65 24 68 70 61 33 11 68 26 21 44 62 58 2 61 13 63 5 62 25 29 21 65 30 35 4 61
36 58 70 62 51 68 53 46 50 6 62 39 2 40 21 44 27 17 15 62 14 19
42 44 19 62 64 65 42 55 76 42 3 69 42 3
39 61 5 35 25 25 79 64 35 51 8 62 71 27 35 67 73 70 62 14 71 1 57 62 54 39 31 40 6 38 28 18 8 37 38 16 65 75 52 63 44 60 46
46 7 61 51 60 50 46 44 24 68 70 6 38 35 12 11 30 64 71 7 8 62 14 64 65 59 27 34 21 71 41 73 55 49 11 49 2 11 69 72
4 70 55 19 72 22 44 61 57 44 40 46 67 76 24 27 3 49 2 72 48 [illegible] 4 19 70 21 51 8 62 65 60 15 70 6 38 6 61 51 32 68 70 65

Step 4: Deciphering with HMMs

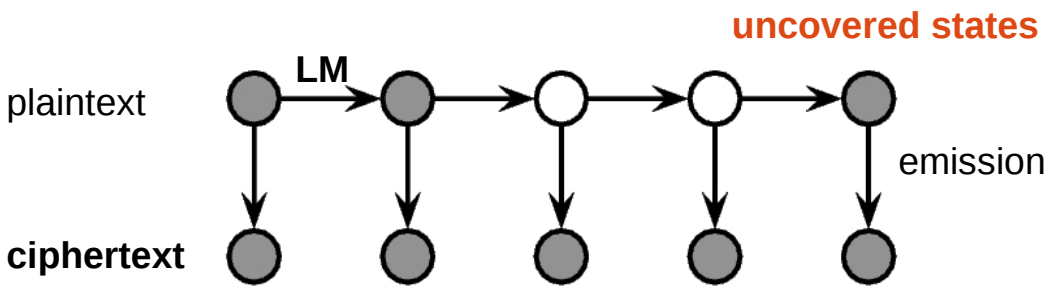
Baseline: K. Knight, A. Nair, N. Rathod, and K. Yamada, "Unsupervised analysis for decipherment problems," ACL, 2006



- Estimate emission probabilities with EM
- Compute Viterbi hypothesis for plaintext sequence

Our approach: Semi-supervised HMM

- Full manual deciphering is laborious
- What about: one line, paragraph, page of a book…?
- Idea: manually uncover some hidden states



- Perform EM and Viterbi on remaining hidden states
- How much ground-truth data do we need?

Increasing complexity of ciphers

- **One-to-one cipher**
  Single cluster for each character

{ن ـنـ ـنـ ن} 3 ن

- **Many-to-one cipher**
  Cluster for each shape of a character

{ـن} 3
{ـنـ} 20
ن

- **Many-to-many cipher**
  Clusters with mix of shapes and characters

{ـن ن} 3 ن
{ـتـ ـنـ} 51 ت

Experimental setup

ALTEC data

- 1 page, 18 lines
- 700 chars, punctuation
- 36+5(+5) target symbols
- 1:1, N:1, N:M

Quran data

- 500 lines
- 42,170 chars, no punctuation
- 36 target symbols
- 1:1, N:1

١ – العرب فى أسبانيا وفرنسا وإيطاليا

خرج العرب المسلمون من الجزيرة العربية فاتحين ، ينشرون دينهم وحضارتهم ، ونجحوا فى فتح جميع اراضى الدولة الفارسية الساسانية ، ثم بدأوا يستولون على اراضى الدولة البيزنطية ، ففتحوا الشام ومصر ، ومضوا فى طريقهم حتى فتحوا شمال افريقيـــة واشرفوا على ســـواحل المحيط الأطلسى . ومضى العرب فى طريقهم يدفعهم ايمانهم الشديد بدينهم ، وحماستهم لعروبتهم ، واعتزازهم بحضارتهم ، وكانت الفتوح العربية كالامواج يدفع بعضها البعض ، فعبر العرب المضيق الذى يفصل افريقية عن اوروبا ، وبدأوا يضعون اقدامهم فى القارة الاوروبية ، وحازوا كثيرا من اقطارها ، ونجحوا فى تثبيت اقدامهم فيها فترة طويلة من الزمن ، نشروا فيها دينهم الرشـيد ، وارســـوا قواعد حضارتهم العربية الزاهرة .

العرب فى اسبانيا :

بدأ العرب اختراق القارة الاوروبية من جهتين ، فحاولوا الدخول اليها من الجنوب الشرقى ففـزوا القسطنطينية عاصمة الدولة البيزنطية ، كما اخترقوها من مدخلهـا فى

٦

Character language model (2-gram, 3-gram)

- MADCAT Phase 1 (news articles)
- 18,096 lines
- 792,764 characters

Building ciphers of different complexity

- **One-to-one cipher**

  Artificial: take plaintext as ciphertext

  ا ي ا ل ا ط ي إ و ا س ن ر ف و ا ي ا ن ا ب س أ ى ف ب ر ع ل ا

- **Many-to-one cipher**

  Artificial: simulate clustering into shapes

  4~ا 3~ي 2~ن 4~ا 3~ب 2~س 1~أ 4~ى 2~ف 1~ب 4~ر 3~ع 2~ل 1~ا

- **Many-to-many cipher**

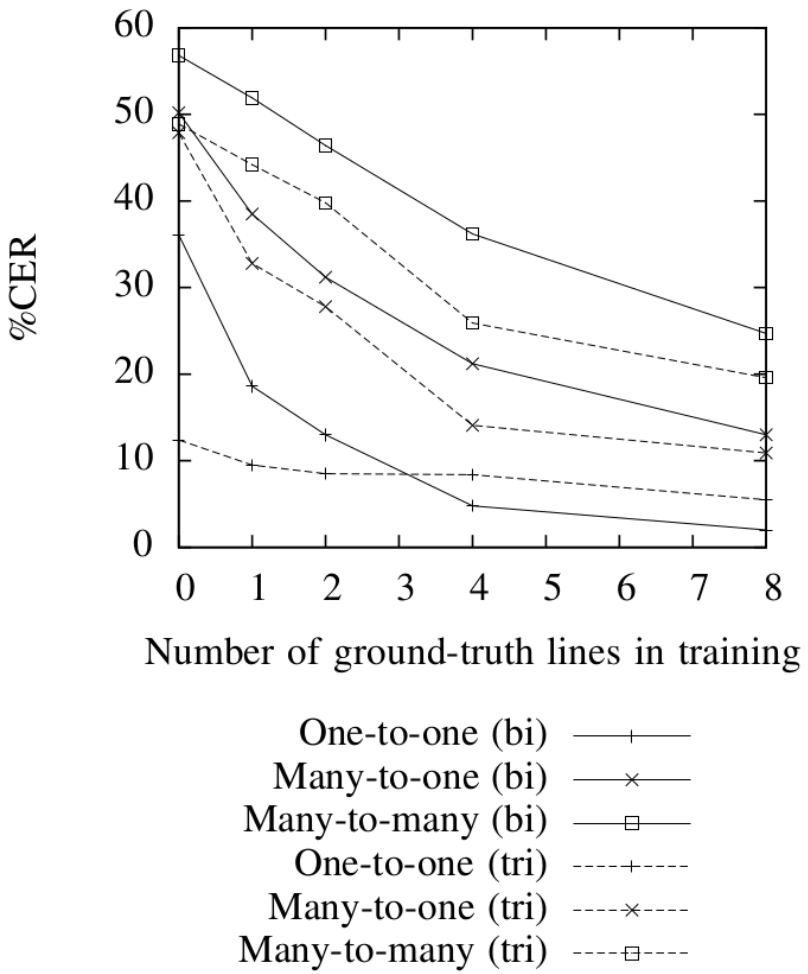
  Outlined OCR approach: manual segmentation, PCA (95%), k-Means (80)

  17 7 19 25 64 65 17 78 76 22 3 76 52 60 65 62 3 55 52

CER on ALTEC data

… after 20 iterations of EM

ALTEC		1:1	N:1	N:M
2-gram	unsupervised	36.1	50.2	56.8
	semi-supervised, 1 line	18.6	38.5	51.9
	semi-supervised, 2 lines	13.0	31.2	46.4
	semi-supervised, 4 lines	4.8	21.2	36.2
	semi-supervised, 8 lines	2.0	13.0	24.7
3-gram	unsupervised	12.4	47.9	48.9
	semi-supervised, 1 line	9.5	32.8	44.2
	semi-supervised, 2 lines	8.5	27.8	39.8
	semi-supervised, 4 lines	8.4	14.1	25.9
	semi-supervised, 8 lines	5.5	10.8	19.6



CER on Quran data

… after 20 iterations of EM

Quran		1:1	N:1
2-gram	unsupervised	19.0	53.3
	semi-supervised, 1 line	17.0	52.0
	semi-supervised, 5 lines	13.1	49.9
	semi-supervised, 10 lines	4.2	46.0
	semi-supervised, 25 lines	2.1	34.6
	semi-supervised, 50 lines	2.1	24.2
	semi-supervised, 100 lines	2.1	15.1
	semi-supervised, 250 lines	0.4	6.2

Still high CER values for many-to-one case, although we have...

- longer ciphertext
- less target symbols

CER on Quran data

… after 20 iterations of EM

Quran		1:1	N:1
2-gram	unsupervised	19.0	53.3
	semi-supervised, 1 line	17.0	52.0
	semi-supervised, 5 lines	13.1	49.9
	semi-supervised, 10 lines	4.2	46.0
	semi-supervised, 25 lines	2.1	34.6
	semi-supervised, 50 lines	2.1	24.2
	semi-supervised, 100 lines	2.1	15.1
	semi-supervised, 250 lines	0.4	6.2

2% of lines

50% of lines

Still high CER values for many-to-one case, although we have...

- longer ciphertext
- less target symbols

Conclusion

- Works well for one-to-one ciphers
- Many-to-one/many ciphers need a lot of ground-truth

Future Work

- Compare with other deciphering approaches
- Higher-order language models, in-domain data
- Assign ground-truth to „hard instances“
- Joint model for clustering and deciphering

Thank you for your attention!