#### **Applications of Subword Spotting**

**Brian Davis** 





PLACE OF ABODE.			NAME	BELATION.	B	E.	PERSONAL DESCRIPTION.				10	CITIZENSHIP.				UCAT	TON.			NATIVITY AND MO	THER TONGUE.			
f.	Lenn	lan,			of each person whose place of abode on January 1, 1920, was in this family.		:	1			4	营	ġ		: 1	1	14	3	3	Place of birth of each per	on and parents of each per	and, in addition, the mother is	atted States, give the state ague. (See Instructions.)	or territory. If of f
ų,	or farm	Ing		-ű,	Inter surtance first, then the stren name and middle holder, if any.	Relation ship of this perma to the head of	H	1			1		ı E		<b>.</b>		1	1.	1.	PERSO	N.	PATE	<b>1</b> .	-
1	-	ar vi	ier of b- ite	and a	Include every parson living on January 1, 1980. Omit children bern since January 1, 1980.		i.		i	-	4	H		37			ļi	11	1	Place of birth.	Hother tengue.	Place of birth.	Hother tangue.	Place of 1
1				•		•	7	8		10	11	13	1	13	14	15	16	17	18	10	30	21	20	33
	V	Γ	Τ		Pond Mary	daughter			7	W	24	¥IS						ges	340	Itali		Utah		2ta
	V	-	-		- Gladus	dalahter			17	W	21	15	T				us	34	Bes	Pitch		Utah		Utal
	V	-			- Generatient	dallation			2	W	11	S				0	un	200	su.	Utah	Sec. 1	Utah		Utas
	1	-	T		- asail J.	Look			in	W	113	115				0	nest	eres	mes	2 utali		Utah		rita
	FM	20	80	41	Kappend George, R.	Kend	R		in	W	33	3m	1			-		20	1200	2itah		rital		Uta
	~	1	1	1	Hagel P.	and.	T		1	w/	34	in	1		1			The	aces	Adaho		2itala		zeta
	V				martin T.	son			m	W	43	45						1.	9	2etah		utah		Ida
	FAA	200	2 01	44	Tall Henry &	head I	O	F	in	W	15%	n	7	-	1			20	200	rital		alrica		alsic
	V	1	T	-1	har had a che	mile		-	2	W	4.5	sin	1					40	lyco	nevada	10.01	Allinois		20
	11		1		muran, X.	and			211	in	2:	25	T					24	leses	ritah	1 1 2 2	ritah		ner
	V	1			- and I.	ion			m	W	18	S	T		1		200	140	244	gitah		2ital)		nevo
	1	1			- Ros alie	daughter.			Ż	W	-115	- 5	T			4	24	80	ine,	2 tale.		2dal		ner
	V	1			- Willow C.	1 mi			ni	W	12	IS	Т		1		res	240	124	Zdal	1.1.1	retal	1	ner
		1	1		- Generieve	daug litre			4	W	10	S	T	-	1	-	3e	Tres	340	Sital.		zitah		nevo
	×	21	0 2	44	Robinson, Elizak	head	R		m	W	35	m	,		1	6		20	ber	Temperdel		Tennesse		Terme
	E.	1	T	-	- Famil	wile			2	W	31	20	1			1		24	Tres	Temessel		Tennessee	-	Tenne
		1			Lester	son			h	W	13	S		-			qui	31	1 40	Tennessee		Termesser		Tenne
		1			- Loravie	daughter.			Ż	W	13	5					m	in an	lues	Fernes all		Jennesse		Tem
	1	1			malem) C.	sol			m	W	11	S	T		1	0	Res	-	here	Tennesses		Tennessee.		Teme
		1			0.1.10				C7		10						1	1	1	1 / 0	1	10		1

PLACE OF ABODE.			NAME	BELATION.	18	E.	PERSONAL DESCRIPTION.				Ţ	CITIZENSBIP.				UCAT	TON.			NATIVITY AND MO	THER TONGUE.			
f.	Lenn	1		-	of each person whose place of abode on January 1, 1920, was in this family.		1	1			4	t	ŝ		;		H	3	3	Place of birth of each per	on and parents of each per	and, in addition, the mother is	sited States, give the state ague. (See Instructions.)	or territory. If of f
ų,	or farm	ing bous	r fai	-ű,	Inter surname first, then the given name and middle holder, if any.	Relationship of this person to the head of	19	1			1				1		1	1.	1.	PERSO	я.	PATE	<b>1</b> .	-
1	-	-	ter of te- ite		Include every parson living on January 1, 1980. Omit children bern since January 1, 1980.		i.		i	-	4	H	Tak I	31	H		ļi	11	1	Place of birth.	Hother tengue.	Place of birth.	Hother tangue.	Place of 1
1				•		•	7	8		10	11	13		13	14	18	16	17	18	10	30	21	20	33
	V	Γ	Τ		Pond Mary	daughter			7	W	24	¥IS						ges	340	Itali		Utah		2ta
	V	1			- Gladus	dalahter			17	W	21	15	T				us	34	Bes	Pitch		Utah		Utal
	V	-	-		Generatient	da litin			2	W	VI	S				0	un	200	su.	Utah	Sec. Sec.	Utah		Utas
	1	-	1		- asail J.	100			in	W	113	15				0	nest	eres	mes	2 utali		Utah		rita
	FM	20	82	41	Kappend George, R.	Head	R		in	W	33	3m	1			-		20	1200	2itah		rital		Uta
	~	1	T	1	- Habel P.	and	1		1	w/	34	in	1		1			The	aces	Adaho		2itala		zeta
	V				montin T.	son			m	W	43	45						1.	9	2etah		utah		Ida
	FAA	200	2 00	44	Tallat Venny 4.	head I	Ø	F	in	W	15%	n	7		1			20	200	rital	1	alrica		alsic
	V	1	T	1	had made C	mile		-	2	W	4.5	sin	1					40	lyco	nevada		Allinois		20
	11		1		muran, 24.	and			211	in	2:	25	Т					24	leses	ritah	1 1 2 2	ritah		ner
	V	1	1		- and I.	ion			m	W	18	S	T		1		200	140	244	gitah		2ital)		nevo
	1	1			- Ros alie.	daughter.			Ż	W	-115	- 5	T			4	24	80	ine,	2 tale.		2dal		ner
	V	1			- Willow C	1001			ni	W	12	IS	T		1		res	240	124	Zdal		retal	1	ner
		1	1		- Generieve	daughte			4	W	10	S	Т		1		3e	Tres	340	Sital.		zital		nevo
	×	24	0 24	44	Robinson, Elizah	head	R		m	W	35	m	,		1	6		20	ber	Temperdel		Tennesse		Terme
	E.	1	T	-	- Formie	wile			2	W	31	20	1					24	Tres	Temessel		Tennessee	-	Tenne
		1			Lester	son			En	W	13	S			1		qui	31	1 40	Tennessee		Termesser		Tenne
		1			_ Loravie	daughter.			Ż	W	13	5					m	in an	lues	Fernes all		Tennesse		Tem
	1				- Malemi C	sof)			m	W	11	S	T		Í	4	Res	-	here	Tennesses	.]	Tennessee.		Teme
		1			0.1.10				C7		10						1	1	1	1 / 0	1	10		1

PLACE OF ABODE.		•	NAME	BELATION.	18	E.	Т	FERR				cr	TISE	SHI		EDU	CAT	ION.	-		NATIVITY AND MO	THER TONGUE.			
f.	Long	1	of Num		of each person whose place of abode on January 1, 1920, was in this family.		:	1		Ι.		4	t	ġ	:	I			3	1	Place of birth of each perm	on and parents of each per	and, in addition, the mother is	sited States, give the state ages. (See Instructions.)	or territory. If of f
Ľ,	or farm	ing bous	r fami	ð,	Bater excesses first, then the given name and middle holder, if any.	Relation ship of this perman to the head of	19	1				1	١,						1.	1.	PERSO	x.	PATE	12.	
11	-	a ri	ter of v b- itation	-	Include every pareon living on January 1, 1980. Omit children bern since January 1, 1980.		ľ		1			-	the second	1				11	11	1	Place of birth.	Hother telgue.	Place of birth.	Hother tongue.	Place of 1
1			•	•		•	7	8	•	1	0	11	13	13	14	1		16	17	18	10	30	21	20	\$3
	V				Pond Mary	daughte			Ŀ	E L	1/2	24	S	Γ			Т		yes	340	Idahi		Utah		2ta
	V	-			la adman	dalation		1	L	- N	1	20	S	1			-	11	41	Res	Pital		Utah		2utal
1	V	-			- la contrart	1. St.		1	12	nu.	rl	1/	S		1		4		200	en.	Zatala)	1.1.1	Utah		2110
	1	+	-		- Chail T	1 augura		1	De la	1 M	1	13	S		1	1	4	n A		200	91tal		Utal		zita
	FAA	20	c 74		V lange R	Kerl	P		12	2014	1	3.3	m	1			1	4	24	-	71tale		Zitale		71ta
	14	1	147	4	The P	inag	"	1	Ľ			24	'n	+	1-	1	-†	-	Le.	and a	Idaho		21tale		717-
		+		ť	in tim	surge.		F	2			3	Ś	+	<u> </u>	1	+	9		The	21tal		21tale		11
	Fre	-	200	Ť	Toll & The and	last	10	r	17			12	m	, —	-		+		2		2.t.h		alica		alia
	FAS	109	127	4	Sallon, Henry F.	nead	¥	r	12	214		10	24	-		1	+	Å	714	-	nan		and	1	217
	1	-		÷	- susannan C	wife	-	1	20	14		20	5	⊢	1		+	a	200	1910	nestada	-	2.7		200
	-7			÷	- myrun x.	son			14			10	5	+-			-	-	Jus .	40	art		2,7 D		24
	1	1.		÷	- annel	eon DL		-	P	100	1		C	+	+		Ż	res ,	44	410	a f. l		2 d		Zu
	-	-		+	- nor give.	daughter,	-	-	1		<u> </u>	5	5	+	+		-ť.	La l	360	34	That	-	2 that	1	200
				÷	- Multon C.	son			P	2 M	ł	2	2	┢			-f	4	yes	yes	ah		utah_		2000
	- V	-		Ť	Of generieve	daug hter	5	-	ť	t n	4	10	5	+	+-		- 12	4	que	yes	21 ale		Tuch		nero
	×	21	0 24	5	Nobinson, Blight	head	R	-	14	M	4	38	m	+	-	-	+	_	7es	que	Jennessee		Timessee		Jenne
				-	- fame	wife			1×	2	Z	31	M	1			-+-		44	yla.	Innersel		1ennessee		Lesone
		-		4	lister	son	$\vdash$	-	14	N	24	3	S	$\vdash$			_	rei .	qui	74	Tennessee		1 emessee		Jenne
		_	_	4	- Loravie	daughter.		_	de la	E L	×4	3	S	1	_		-6	14	94	44	Tennesses.		Jennessee		Jem
				-	- Malgom C.	son	_		14	1	24	11	S	1	_		12	Kel .	yei	yes	Tennessee		Temessee		Terme
1		1		- 1	A. (.11)	and the second state of th		1	L.	1	. 1	0	-	1								1		1	

'n	PLACE OF ABODE.		ODE.	NAME	BELATION.	TERME.	PERSONAL DESCRIPTION.	CITIZENSHIP.	EDUCATION.	NATIVITY AND MOTHER TONGUE.					
1111		A STATE		of each person whose piece of abode on January 1, 1920, was in this family. Beter senses first, then the stress name and middle builds from her days in fact. Include every person Bring on January 1, 1900. Omit	Belathen ship of this permit is the band of the family.	and from a	er or mon. a lant berb- day. author	a traditor		Place of birth of each person and parents of each p PEBSON. Place of birth.	Place of birth.	ted States, give the state gas. (See Instructions.) 	er territery. If of f		
1			•	•	•	Mc	huldn't it	ho nio	d' if' co	mothing could	21	20	\$3		
	V			Pond Mary	daughty	VVC	Julun Lit	De IIIC		menning could	Utah		21ta		
	V	-	_	- Gladys	daughter		scan a	automa	atically	for you?	Utah Utah		uta.		
	1	-	1	- asail J.	Lon	T .	WW 1315	4	R. F. E. B. D. C.	Tutan	Utah		uta		
	FM	20	\$ 243	Karren George R.	Head !	R	20 W 33 m		340 200	2itah	rital		Uta		
	~	1		- Hazel P.	unte		7 W 34 m		44 9100	Idaho	2italy		zeta		
	V			- montin J.	son		m W 43 S		9.9	2etah	utah		Ida		
	FM	209	244	Tallat. Henry S.	head	ØF	WW 57 m		nes no	ritak	africa		alric		
	V	Γ.		- Susannaly C	will		2W48m		Ages yes	nevada	Allinois		20		
	15		1	- muran X.	son		20/12/22/5		24 24	2itah	2 tak		nern		
	V			- annel I.	eon		mNISS		Thes Thes 2410	zetah	ritah		new		
	1	1	1.1	- Rosalie	daughter.		2 11/5-5		34 30 30	ritaly.	2dal	1	ner		
	V			- Hilborn C.	son		21 W12 S		tay yes yes	2dah	utah		ner		
	. V			- Genevieve	daughter.		5 N/10 S		The yes yes	Zitah.	zitah		nero		
	×	210	245	Robinson Elizah	head 1	R	m W 38 m		3 res sus	Termensee	Tennessee	-	Term		
				- Famil	wife		Z. W 31 M		34 44	Tennessee	Tennessee	-	Terme		
				lister	son		M W 15 5		yes yes yes	Tennessee	Termessee		Tenne		
	1			Lovavie	daughter.		ZW135		see yes yes	Fennesses.	Tennessee		Tem		
				- Malgom E	son		m W/1/5		345 yes yes	Tennessee	Tennessee		Teme		
	1000	1		A.C.10											

We're going to do this and other things with subword spotting

## Outline

- Review word spotting
- Subword spotting
  - Our implementation
  - Performance
- Applications
  - Suffix spotting
  - Transcription assistant demo

## Word Spotting

- Goal is to search corpus of images directly
- Query-by-string (QbS): search with text
- Query-by-example (QbE): search with an example word image

Search for "**pay**"

Search for "payment"

as paid, and how - as paid, and how d Sugeants pay, i d Sugeants pay, i ext payment: as ext payment: as

## Subword Spotting

- We now allow spottings within words

Search for "**pay**" Search for "pa" as paid, and how as paid, and how -& Sugeants pay, i d Sergeants pay, a ext payment: as ext payment: as

## Subword Spotting Implementation

- Converted Sudholt et al's word spotting method PHOCNet to perform sliding window
- Changed PHOC used and comparison method for better QbS results
  - Less resolution for PHOC
  - Similarity based on cross-entropy instead cosine distance
  - Original PHOC and cosine similarity better for QbE
- Found optimal window width for each subword of interest



#### Datasets

Bentham US 1930 Census Names formad by the reconsideration of all endence Cumungham, John H. Corrigan George B delivered in the course of the original exami Edelin John N iow Letitia J. by sichness or other cause from officating in the exercise of his Griffith John B duty and a Separty or substitute as will then be necessary Voll. Murry C: O'Hana, Charles P. 1 CX 1 in Manne 5. Those of a Guardian over his Ward. See Law of Guard. Funch Catherine a and and Wards. Got mitter alian 4

## Subord Spotting Results

- Unigrams: all letters of alphabet
- Bigrams: 100 most frequent in English
- Trigrams: 300 most frequent in English

Reported as Mean Average Precision

#### Bentham US 1930 Census Names

	Unigrams	Bigrams	Trigrams	Unigrams	Bigrams	Trigrams
QbS	67.7%	68.2%	70.5%	49.7%	40.2%	36.3%
QbE	51.1%	56.9%	57.1%	34.0%	29.5%	28.5%





int





## A searching task

What if I wanted to find all the towns in a set of German documents?

- What about automatically finding all words ending in "-burg"?

Nr. 7. Give bierry an grang and grang weging form Bor bem unterzeichneten Stanbesbeamten erschienen beute zum Bwede ber Chefchließung: 1. ber Buskallung Bund Growner Rersten ber Persönlichteit nach Sury drin yan fawfore between ban Buifbuirns adrew Gibiegow numer tant, noncegalipper Religion, geboren ben fierfand ponenjeg ; At a Francisco \_\_ bes Jahres taujend acht hundert ferifying mind dary : Suban , wohnbaft in Minblesden Con vel akundingnan Jeferen Jattlind Therstern mid verfor Gafrian Grighein for Tweather woodstat 311 Suban, froging Traffor 2. vie Carting Conformier Brikin your man Eindusy ber Perfonlichteit nad lon fannt.

Pinni

# Suffix Spotting

- Find all words with a given suffix
- Constraint on original subword spotting problem
- Could be extended to handle regular-expression-like queries



# Suffix Spotting

- IAM results



# Suffix Spotting

- Census Names results



AP

- Using ground truth word segmentations
- Both embedding and PHOCs for windows are precomputed.
  - Selection snapped to closest window
  - ~10 second delay to compute PHOC for all windows of a single size, depending on size, using GPU

# Thank you

Questions?

(Images in case of technical difficulties)



(Images in case of technical difficulties)



(Images in case of technical difficulties)



## Subord Spotting Implementation

- Network architecture

