# An Interface for Assisted Curation of Knowledge Bases from Unstructured Text

Jeremy Wright Computer Information Systems University of the Fraser Valley Abbotsford, Canada Email: jeremy.wright@student.ufv.ca Gabriel Murray Computer Information Systems University of the Fraser Valley Abbotsford, Canada Email: gabriel.murray@ufv.ca Ben Hachey School of Information Technologies University of Sydney Sydney, Australia Email: ben.hachey@sydney.edu.au

Abstract—We present an interactive graphical tool for assisted curation of knowledge bases from unstructured text data. Given text input, the user can create a knowledge base from scratch, including sub-tasks of entity mention annotation, matching mentions that refer to the same entity, and extracting relations between entities. The interface is designed to enable organizations to extract valuable knowledge from text data that may otherwise remain unexploited. We evaluate the interface through a formative user study. The study results suggest several key directions for refinement. Results also highlight the efficacy of the interface: all participants were able to create a knowledge base from scratch.

*Keywords*-knowledge base population; entity matching; relation extraction; intelligent interfaces; natural language processing

### I. INTRODUCTION

Many organizations possess vast amounts of textual data that is not exploited for decision-making and business intelligence purposes because of its unstructured nature. We create an interactive graphical tool that can facilitate the creation of a knowledge base (KB) from such unstructured textual input, a task known as knowledge base population (KBP). Consider the following example text:

> Barack Hussein Obama II is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

> *Obama* was born on August 4, 1961, at *Kapi'olani Maternity & Gynecological Hospital* (now *Kapi'olani Medical Center for Women and Children*) in *Honolulu, Hawaii*, and would become the first President to have been born in *Hawaii*.

The tool allows the user to identify mentions of entities (e.g., *Barack Hussein Obama II*, *Honolulu, Hawaii*, *Illinois Senate*), label the type of each entity (e.g., person, location, organization), determine when multiple mentions are referring to the same entity (e.g., *Barack Hussein Obama II* and *Obama*), identify relations between entities (e.g., *Barack Obama* was born in *Honolulu, Hawaii*), record offsets for entities, and enter entity descriptions (e.g., *Barack Hussein Obama II* is the leader of the U.S.A.).

Assisted manual creation of a knowledge base is particularly valuable, as fully automated techniques for these sub-tasks still suffer from low accuracy [1]. Studies of assisted curation in the biomedical literature demonstrate that it can improve throughput [2], but suggest that user experience should drive design of both user interface and automated KBP techniques [3]. Previous work discusses design principles for annotation tools [4]. However, to our knowledge, no previous work explicitly evaluates interface design for such KBP annotation tools.

The main contributions of this work are as follows:

- We present an interactive graphical interface for creating a knowledge base from unstructured text data.
- We validate the KBP curation tool through a formative user study.
- We are making the software and source code publicly available upon publication.

The formative user study consists of a self-guided cognitive walkthrough in which participants were asked to complete eight tasks. The results are encouraging in that all users were able to successfully create and save a knowledge base from text input. Key areas for improvement include providing better guidance and more explicit feedback. Participants familiar with NLP found the tool particularly easy to use.

The structure of the paper is as follows. Section II describes related work on knowledge base population and document-level annotation. Section III describes our interactive tool and provides several sample screenshots. Section IV describes the structure of the user study, and the results are presented in Section V. In Section VI we describe our planned future work and extension of the interactive tool. Conclusions are presented in Section VII.

https://en.wikipedia.org/wiki/Barack\_Obama

## II. RELATED WORK

Knowledge base population: KBP builds a repository of formal knowledge by extracting information from text collections [5]. The resulting KB can be used for semantic queries, e.g., retrieving Honolulu, Hawaii for the query "Where was Barack Obama born?". The US National Institute for Standards and Technology has run shared tasks on KBP as part of the Text Analysis Conference (TAC) since 2009. The primary tasks have been entity discovery and linking (EDL) [6], and slot filling (SF) [1]. EDL has evaluated query-driven and document-level named entity handling identifying mentions of entities in text, and clustering the mentions that refer to the same entity (e.g. "Barack Obama" and "President Obama"). SF has evaluated query-driven and collection-level relation extraction - identifying all possible facts for a given entity (e.g. the fact that Barack Obama is president of the United States). EDL and SF can be combined for end-to-end KBP.

KBP is part of the larger field of natural language processing (NLP), which aims to enable computers to analyze, represent and generate human languages (as opposed to programming languages or formal mathematical languages). In turn, NLP is part of the larger field of Human-Computer Interaction (HCI), where the medium of interaction is natural human language.

*KBP for retrieval:* Guo et al. [7] calculate that 71% of web search queries include entities. In addition to building a formal knowledge repository, KBP is also a means for semantic indexing of documents. It allows users to retrieve documents that mention specific entities instead of relying on keyword searches based on ambiguous name strings. Users can review evidence supporting facts in the knowledge base. And, with the right tools, users can correct facts.

Interactive KBP: Research in KBP to date focuses on fully automated techniques [1]. However, accuracies are not high enough to deploy in most settings. In the TAC slot-filling task, for example, inter-annotator agreement is 70% F-score. The best system uses active learning to source additional human annotation, but still scores 37% [8]. This motivates interactive approaches that incorporate intelligent algorithms to streamline knowledge curation with human accuracy. Text mining is used in document curation for realworld biomedical databases [9] and elsewhere [10]. Previous work demonstrates that this can improve throughput for annotation of genes, alleles, and relations [2], selection of documents for curation [11]; and gene indexing [12]. But [3], [13] caution that user experience should drive design of automated KBP techniques as well as the user interface.

*Large-scale KBP:* In recent years, systems have exploited massive amounts of data to extract more facts with confidence [14]–[16]. Google reports very high coverage and precision when using very large web data [17]. However, these approaches do not generalise well to smaller document

collections that cannot exploit redundancy at the same scale. In interactive KBP, an organization can control the tradeoff between cost and coverage. They can leverage algorithms to inform annotation of rare facts, rather than ignoring them.

Annotation tool design: There are a number of tools for document-level annotation in support of KBP [18]–[20]. Alex et al. [3] briefly summarise interface design issues for assisted curation of protein interactions in biomedical research papers: "what information is displayed to the curator, in what form, and what kind of manipulations can the curator carry out?" Pontus et al. [4] state design principles for the Brat annotation tool:

We believe that intuitive and user-friendly interfaces as well as the judicious application of NLP technology to support, not supplant, human judgements can help maintain the quality of annotations, make annotation more accessible to non-technical users such as subject domain experts, and improve annotation productivity, thus reducing both the human and financial cost of annotation.

However, to our knowledge, no previous work evaluates interaction design or user experience for KBP annotation.

## III. DESCRIPTION OF THE INTERACTIVE TOOL

The knowledge management tool presented here has a number of features that make the extraction and management of unstructured text documents easier. The main goal of this interactive tool is to make knowledge base curation as quick and easy as possible by providing intuitive and simple features that will reduce the learning curve significantly. Another objective is to limit human error as much as possible.

There are four important areas of this interface as seen in Figure 1: The "Mentions" tree, the "Entities" tree, the text viewing area and the button bar. The "Mentions" tree is where all annotations from the text are stored for easy access. The tree has four columns which represent the important parts of a mention: ID, Name, Class, and Description. The ID is a unique key for each mention. The numbering starts at 1001 and increments every time a new mention is created. The Name is the non-unique name of the mention; in some cases, there could be multiple mentions of the same entity, and they can all be uniquely stored. The class column contains the class information of the mention. Class can be either PER, for a person, GPE for a geo-political entity, or an ORG for an organization. The last column holds the description, which is unrestricted text provided by a user to give a meaningful description of the entity. Each mention also gets a character offset within the text. This offset information is stored in a dictionary related to each unique key, and when the user clicks on a mention, the text area shows where the mention is from, by highlighting the entity within the document. At the same time, the "Entities" tree will highlight the entity mention matches.

76 ColdStart GUI 1.	2.1	And in case of the local division of the loc	-	and the Party of the local division of the l				Contraction of the second se								
File									Info							
- 🖧 🛛 🗮	📄 📄 🏜															
New Save					Match I	Relate	Relate	Edit Delete Helj	p Quit							
Mention II	)'s	lame	Class	Description	C:/Users/	/JWrigh	t/My Sch	chool/NLP Work-Study/DBCcoperWikiTestDoc.txt								
M:1003	hijacked		PER		C:/Users/	JWrigh	t/My Sch	<pre>chool/NLP Work-Study/ColdStartGUI-1.0.10.py</pre>								
M:1023	Portland S	ate University	ORG	University		_										
M:1022	Geologist	eonard Palme	r PER	Scientist	-	at .										
M:1021	Himmelsb	ach	PER	Scientist	New	Doc 1	Doc 2	2 Doc 3	Dump							
₩ see M:1004	Boeing		ORG													
org:member	s M:1017		ORG		of the Ho	obbs àc	n a Port	ttiend grand jury returned an indictment against "John Doe, aka Dan Cooper", in absentia for air pirady and vi The indictment in effect formally initiated prosecution of the bijacker that can be continued, should be be	oracion -							
M:1005	Portland, 0	regon	GPE		apprehended, at any time in the future. [64]											
M:1006	Seattle, Wa	shington	GPE													
- M:1007	FBI		ORG		Physical To 1070	eviden	ce									
org member	s M:1013		ORG		miles (2)	l km) e	ast of C	caring instructions for lowering the alt starts of a /2/ was found by a deer huncer her a logging road about Castle Rock, Washington, well north of Lake Merwin, but within the basic path of Flight 305,[67]	13							
M:1008	American		GPE													
use M:1009	Boeing 727	aircraft	ORG			-										
use M:1010 aircraft			PER		Portion o	Portion of Backet Hypen 3 1980 Giscovery In February 1980 an eight-vest-old box named Brian Ingram, vacationing with his family on the Columbia River about 9 miles (14 km)										
Entity ID's	Mamo	1	Claure	Description	downstrea	am from	Vancouv	yer, Washington, and 20 miles (32 km) southwest of Ariel, uncovered three packets of the ransom cash, signifi	cantly							
Entry ID'S	D.D.C.	Class	Description	disintega	disintegrated but still bundled in rubber bands, as he raked the sandy riverbank to build a campfire.[68] FBI technicians confirmed that											
₩ E:1001	D. B. Cooper		'EK		the money	y was 1	ndeed a	a portion of the ransom-two packets of 100 twenty-dollar bills each, and a third packet of 90, all arranged in	the							
ME1001					Same orde	er as w	nen give	ver to cooper.[ca][/o]								
NE1011	to the second		0.00		The disco	overy 1	aunched	i several new rounds of conjecture, and ultimately raised many more questions than it answered. Initial states	ments by							
E1002	Region		PER DRG		investiga	ators a	nd scien	entific consultants were founded on the assumption that the bundled bills washed freely into the Columbia Rive	r from							
M41004	overng		5110		one of it	and we	connect	thing tributaries. An Army Corps of Engineers hydrologist noted that the bills had disintegrated in a "rounded	·*							
M-1009					buried. [7	71] If	so, it c	confirmed that Cooper had not landed mean Lake Merwin, nor in any other part of the Lewis River, which feeds	into the							
M-1010					Columbia	well d	ownstream	eam from the discovery site; and it lent credence to supplemental speculation (see Later developments above) p	lacing							
M-1016					the drop	zone n	ear the	Washougal River, which merges with the Columbia upstream from the discovery site.[72]								
D F-1020	Portland State University	ity (	DRG U	niversity	But the	tree t	loating"	" hypothesis presented its own difficulties: it did not explain the ten bills missing from one packet, nor wa	a there							
D E-1019	Geologist Leopard Pal	mer F	DER S	rientist	a logical	l reaso	n that t	the three packets would have remained together after separating from the rest of the money. Physical evidence	was							
D E:1018	Himmelshach PER Scientist			cientist	incompatible with geologic evidence: Himmelsbach observed that free-floating bundles would have had to wash up on the bank "within a couple											
▷ E:1004	Portland, Oregon G		SPE		of years" of the hijacking: otherwise the rubber bands would have long since deterioreted, [73] an observation confirmed experimental the Cooper Research Team (see Ongoing investigation below).[74] Geologic evidence suggested, however, that the bills arrived at the their discovery-a beach front known as Tina (or Tena) Bar-well after 1974, the year of a Corps of Engineers dredging operation on th											
T E:1005	Seattle, Washington G		SPE													
M:1006					stretch o	of the	river. G	Geologist Leonard Palmer of Portland State University found two distinct layers of sand and sediment between	the clay							
▷ E:1006	FBI ORG			deposited on the river bank by the dredge and the sand layer in which the bills were buried, indicating that the bills arrived long after												
▷ E:1007	American GPE				were natu	were natural decodests. That finding, if true, favors an arrival time of less than one year after the event (based on the rubber band										
▷ E:1009	Special Agent Larry Carr PER					nt), bu	t does n	not help to explain how the bundles got to Tina Bar, or from where they came [76]								
⇒ E:1010	Portland International	Airport C	DRG													
🗹 Tool-tips					Muttiple	altern	arive rh	chaories were advanced tome surmised that the money had been found at a distant location by someone (or noss) Search	nuy even							

Figure 1. This image shows the four main sections, the "Mentions" tree, the "Entities" tree, the text viewing area and the button bar.

All colspan="2">OPE   OPE	76 ColdStart GUI 1.	2.1	a state of	State of Street, or other	States I and special vol.	-	THE R. R. LANSING MICH. CO., LANSING MICH. & LANSING MICH.	-	-	-					
New   Load   Lead   PER   ORG   OF   Match   Redate   Redate <thredate< th="">   Redate   Redate<th>File</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th>Info</th></thredate<>	File											Info			
Member   Name   Case   Description     M1005   Portland, Oregon   GFE     M1006   Seattle, Wendyndown   GFE     M1007   FRI   OR5     orgenmebers   M1013   OR6     M1008   American   GFE     M1009   American   GFE     M1000   anicraft   FRI     M1010   anicraft   FRI     M1010   anicraft   FRI     M1010   anicraft   FRI     M1011   Cooper   FRI     M1012   the agency   OR6     M1013   Due cooper   FRI     M1014   Portland Intermetionality   OR6     M1015   Filed M105   OR6     M1010   Due cooper   FRI     M1011   Portland Intermetionality   OR6     M1012   Due agency   OR6     M1013   Due cooper   FRI     M1014   Portland Intermetionality   OR6     M1015   Filed M235   OR6     M1010   Due cooper   FRI <td< th=""><th>new Save</th><th>Load Load PE</th><th>R OF</th><th>k 📀 RG GPE</th><th>og 📑 Match Relate</th><th>ج Relate</th><th></th><th>/ Edit</th><th>Delete</th><th>Ca Delete</th><th>🕑 Help</th><th><b>ui</b> Quit</th></td<>	new Save	Load Load PE	R OF	k 📀 RG GPE	og 📑 Match Relate	ج Relate		/ Edit	Delete	Ca Delete	🕑 Help	<b>ui</b> Quit			
MADDS   Portland, Gregon   GPE     MADDS   Sastett, Washington   GPE     MADDS   Sastett, Washington   GPE     MADDS   MADDS   Opc   Doc   Doc   Durp     MADDS   MADDS   GPE   Doc   Durp     MADDS   MADDS   GPE   Doc   Doc   Durp     MADDS   MADDS   GPE   Doc   Doc   Durp     MADDS   Second   GPE   Doc   Doc   Durp   Durp     MADDS   Second   GPE   Doc   Doc   Durp   Durp   Durp     MADDS   Second   GPE   Doc   Doc   Durp   Durp   Durp     MADDS   Second   GPE   Doc   Doc   Durp   Durp   Durp     MADDS   Dec   Doc   Doc   Doc	Mention I	D's Name	Clas	ss Description	C:/Users/JWri	ht/My School	1/NLP Work-Study/DBCooperWikiTestDoc.txt								
MAX000   Seate, Washington   GPE   OPE	M:1005	Portland, Oregon	GPE		CT/OSEIS/OWII	inc/ky sendo.	T/REF WORK-Study/coldstartSol-1.0.10.py		_	_	1	_			
	M:1006	Seattle, Washington	GPE												
MADD2   Annican   OPG     MADD2   Annican   OPG     MADD2   Beeng 727 wirett   OPG     MADD3   Beeng 727 wirett   OPG     MADD4   OPG   OPG     MADD3   Seeng 727 wirett   OPG     MADD3   Seeng 727 wirett   OPG     MADD4   Cooper   PR     MADD3   Seenal Apentiany Cooper   PR     MADD3   Descendar Cooper   PR     MADD4   Cooper   PR     MADD5   Descendar Cooper   PR     MADD4   Pottad Ametianetional Arr OPG   OPG     MADD5   Descendar Cooper   PR     MADD5   Descendar Cooper   PR     MADD5   Descendar Cooper   PR     MADD5   Descendar Cooper   PR     DiAD15   Descendar   Descendar     DiAD15   Descendar   Descendar   Cooper     DiAD16   Filed Songer   PR   Descendar   OPG     DiAD16   Mado4   PR   Cooper   PR   Descendar   Descendar   Descendar   <	→ M:1007	FBI	ORG		New Doc 1										
Mixed Description   Mixed PR   Mixed Prime   Mixed	org:member	ns M:1013	ORG								1				
MADIO   scored   PR     MADIO   Cocyce   PR     MADII   the signcy   OG     MADII   packad hermitonalAry Con   PR     MADII   Data Cooper   PR     MADII   Data Cooper   PR     MADII   Data Cooper   PR     MADII   Data Cooper   PR     DAMDII   Data Cooper   PR     DAMDII   Data Cooper   PR     DAMDII   Data Cooper   PR     DAMDII   Data Cooper   PR     Directory   PR   Cool Accept   Cool Cooper     MADII   Decomptoint   Cool Cooper   PR     Directory   Name   Decomptoint   Cooper (16)(70)     MADII   Decomptoint   PR   Decomptoint   Cooper (16)(70)     MADII   Decomptoint   PR   Decomptoint   Decomptoint   Decomptoint   Decomptoint </td <td>M:1009</td> <td>Roeing 727 aircraft</td> <td>ORG</td> <td></td> <td>prosecutor com</td> <td>ild argue the</td> <td>at Cooper had forfeited immunity on any of several valid technical grounds.</td> <td>[64][65]</td> <td>The ques</td> <td>stion was</td> <td>rendered</td> <td>moot _</td>	M:1009	Roeing 727 aircraft	ORG		prosecutor com	ild argue the	at Cooper had forfeited immunity on any of several valid technical grounds.	[64][65]	The ques	stion was	rendered	moot _			
MM01   Cooper   PER     MM012   the spinop   066     MM013   Special Agentiany Car   PER     MM014   Pothand Identificationy Car   PER     MM015   Data Coper   PER     MM016   Data Coper   PER     MM017   Name   Class   Description     MM018   D.B. Coper   PER     MM019   D.B. Coper   PER     MM010   MM010   PER     MM010   MM010   PER     MM010   MM010   PER     MM010   MM020   PER     MM020   MM020   PER     MM020   MM020   PER     MM020   PER   The discovery launched several new conder ind scientific consultant sever conduct on the assence relatione core severa mactod core severa mactod core severa mactod core seve	M:1010	aircraft	PER		of the Hobos	ct.[66] The	indictment in effect formally initiated prosecution of the hijacker that o	an be co	ontinued,	should he	be	-4011			
Mill   the spency   ORG   Physical     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of a 127 was found by a deep hunter near a logging road about 13     Mill   Potender diremational start of the start of a 127 was found by a deep hunter fithe start of a 127 was found by a deep hunte	M:1011	Cooper	PER		apprehended.	t any time	in the future. [64]								
b M1013 Special Agent Luny, Carr PER M1014 Portand Plantementional Agr (DR S M1014 Portand Plantementional Agr (DR S M1015 Dua Cooper PER M1015 Dua Cooper PER M1015 Dua Cooper PER M1016 Filight 305 OBS Portand State University of the Second Planter Agent and State University of the Second Planter Agent and State University OBS University Planter Agent and State University OBS University Planter Agent and State University OBS University Planter PER Sciential Humid Res State University OBS University Planter PER Sciential Humid Res State University OBS University Planter PER Sciential Humid Res State University OBS University Planter PER Sciential Humid Res State Planter Agent Age	M:1012	the agency	ORG		Physical 74 C	oldStart GUI 1.2.1									
Mild   Pertiand intermitional Mild   Name   Obs   Name   Class   Description     Easty MD1   Name   Class   Description   Image: Class   Description     Easty MD1   D.B. Cooper   PER   One Class   Description   Image: Class   Description     Mild1   D.B. Cooper   PER   Description   Image: Class   Description   Image: Class   Description     Mild1   D.B. Cooper   PER   Description   Description   Description   Description   Description     Mild1   D.B. Cooper   PER   Description   Descri	▷ M:1013	Special Agent Larry C	arr PER		In 1978 a	In 1979 Choose entity to link. the aft stairs of a 727 was found by a deer hunter near a logging road about 13									
p Mails Und Coper Per p Mails Und Coper Per p Mails Und Coper Per tenty Dro. Name Cose Description Construction of the rankom-two packets of lot twenty-dollar bills each, and a third packet of 90, all arranged in the the glace per Per tenty Dro. Coper PER b Eloo b specked PER p Eloo b Solution of the rankom-two packets of 100 twenty-dollar bills each, and a third packet of 90, all arranged in the meet often at when glaven to Cooper. [69](70] The discovery launched several new rounds of conjecture, and ultimately raised many more questions than it answered. Initial statements by investigators and accentific computations were founded on the assumption that the bundled bills washed freely into the Columbia River from Mailo b b Eloo b Polland State Univerity b Eloo B Polland State Univerity D Eloo B	M:1014	Portland Internation:	al Airp ORG		miles (21 Ma	008 💳 🛛 gperre	sidents_of_country - M1011 - or Lake Merwin, but within the basic path	of riign	10 305.[6]	4					
Prime     One     One     Operation     Operatinstheadd from the stand framode an	D MILULD Dan Cooper PER														
Entry Drs     Name     Clas     Description       0 we E1001     Description     Issuessmean cash, significantly issuessmean cash, significantly Millot       Millot     PE     Second     Description       Millot     Biolog     Description     Description       Millot     Biolog     Description     Description       Millot     Biolog     Description     Description       0     Description     Description     Description <td< td=""><td>P WILDIO</td><td>Tingit 505</td><td>ONO</td><td>T</td><td>Portion c</td><td>C Accept</td><td>Cancel</td><td>in Pirma</td><td>about 0</td><td>milon (14</td><td>lem )</td><td></td></td<>	P WILDIO	Tingit 505	ONO	T	Portion c	C Accept	Cancel	in Pirma	about 0	milon (14	lem )				
C we Hold D.B. Cooper PER disintegrated but still bundled in rubber bands, as he raked the sandy riverbank to build a campfire (68) FBI technicians confirmed that the money was indeed a portion of the ransem-two packets of 100 twenty-dollar bills each, and a third packet of 90, all arranged in the money was indeed a portion of the ransem-two packets of 100 twenty-dollar bills each, and a third packet of 90, all arranged in the money was indeed a portion of the ransem-two packets of 100 twenty-dollar bills each, and a third packet of 90, all arranged in the money was indeed a portion of the ransem-two packets of 100 twenty-dollar bills each, and a third packet of 90, all arranged in the money was indeed a portion of the ransem-two packets of 100 twenty-dollar bills each, and a third packet of 90, all arranged in the money was indeed a portion of the ransem coder as when given to Cooper. (49)(70)   billob billob The discovery launched several new rounds of conjecture, and ultimately raised many more questions than it answered. Initial statements by investigators and acientific consultants were founded on the asymption that the bundled bills washed freely into the Columbia River from the discovery like and in, any goperad to having been dollar bills dollar bills dollar bills and disintegrated in a "rounded" failed and the rais to cooper. (40) the rais to the packet of any photoe to having been dollar bills the rais to cooper. (40) the rais to cooper the rais to cooper. The rais to cooper. The rais to cooper. The rais to coope	Entity ID's	Name	Class	Description	downstream fr	m Vancouver	, Washington, and 20 miles (32 km) southwest of Ariel, uncovered three pack	ets of t	the ransor	n cash, si	gnificant	tly			
Mil01   Mil02     Mil01   mil01     Mil01   mil01     Mil01   mil01     Mil02   highestal     PER   Find State given to Cooper. [63][70]     Mil03   mil01     Mil04   me cader as when given to Cooper. [63][70]     Mil05   me cader as when given to Cooper. [63][70]     Mil06   me cader as when given to Cooper. [63][70]     Mil07   me cader as when given to Cooper. [63][70]     Mil08   me cader as when given to Cooper. [63][70]     Mil09   me cader as when given to Cooper. [63][70]     Mil09   me cader as when given to Cooper. [63][70]     Mil09   me cader as when given to Cooper. [63][70]     Mil09   final set match cooper lad sectific consultants were founded on the asymption that the bills had disintegrated in a "counded" final set miles and consultants were founded in the line set match cooper lad not landed near Lake Mervin, nor in any other part of the Levis River, which feeds into the discovery site. [72]     Mil00   fills discovery asymption that the Columbia well dowardsem from the discovery site. [72]     bill09   followardsem from the discovery site. [72]     bill09   Geologat Loward PRF   Somit     bill010   the "feer flowards when dowardsem fr		D. B. Cooper	PER		disintegrated	but still b	undled in rubber bands, as he raked the sandy riverbank to build a campfire	.[68] FE	I technic	cians conf	irmed the	at			
b E000 Name PER   0.1001 Samp OHG   0.1002 Samp OHG   0.1003 Samp OHG   0.1004 Samp OHG   0.1005 Samp OHG   0.1004 Samp OHG   0.1005 OHG OHG   0.1004 Samp OHG   0.1005 OHG OHG   0.1004 OHG OHG   0.1005 OHG OHG   0.1004 OHG OHG   0.1005 OHG OHG   0.1004 OHG OHG   0.1004	M:1001				same order as	when given	to Cooper.[69][70]	u packet	, OI 90, 6	iii arrang	ed in the	-			
State     ONG     The discovery launched several new rounded on the asymption that the bundled bills washed freely into the Council a statements by investigators and acientific consultants were founded on the asymption that the bundled bills washed freely into the Council a statements by investigators and acientific consultants were founded on the asymption that the bundled bills washed freely into the Council a statements by one of its many connecting tributaries. An Anny Corps of Explosers hydrologist noted that the bills had disintegrated in a "counced" within the object with the founded bills washed freely into the Council and the council and the counce of the state distributaries. An Anny Corps of Explosers hydrologist noted that the state distributaries into the council and the counce of the state distributaries above) placing the drop zone near the Mashougal River, which merges with the Council a state discovery site; [72]       b E1000     Potland State University     Description of the conserve site; and it is on explore the the bills missing from one packet, nor was there a logical reason that the the people vould have remained cogether after separating from the rest of the noney. Physical evidence was a longmatic reserve sould have remained cogether after separating from the verb do the week had to washe who on the bank within a couple	▷ E:1002	hijacked	PER							2.000					
Mildle     one of its many connecting tributaries. An Army Corps of Engineer's hydrologist noted that the bills had disintegrated in a "counded" Mildle       Mildle     Safeing, and vers matted together, indicating that they had been deposited by tiver sation, as opposed to having been diluterately builds. (Till Fas, it confirmed that Cooper had not landed near Lake Mervin, nor in any other part of the Levis River, which feeds into the builds. (Till Fas, it confirmed that Cooper had not landed near Lake Mervin, nor in any other part of the Levis River, which feeds into the builds. (Till Fas, it confirmed that Cooper had not landed near Lake Mervin, nor in any other part of the Levis River, which feeds into the builds. (Till Fas, it confirmed that Cooper had not landed near Lake Mervin, nor in any other part of the Levis River, which feeds into the builds. (Till Fas, it confirmed that Cooper had not landed near Lake Mervin, nor in any other part of the Levis River, which feeds into the builds. (Till Fas, it confirmed that Cooper had not landed near Lake Mervin, nor in any other part of the Levis River, which feeds the drop zone near the Mashough River, which merges with the Columbia upstream from the discovery zite. (T2)       b El03     Geologiat Lowed Fame     PER Somitat a logical reason that the there packts would have remained coopether after explain the test of the money. Physical evidence was lowed a logical reason that the deelogic evidence: Rimeelbach Observed that there-floated physical would not be not of the noney. Physical evidence was lowed a logical reason that the deelogic evidence: Rimeelbach Observed that there-floated physical would have we had to wash up on the bank of would have had to wash up on the bank of the solution.		Boeing	ORG		investigators	and scienti	veral new rounds of conjecture, and ultimately raised many more questions t fic consultants were founded on the assumption that the bundled bills washe	d freely	inswered.	Columbia	River fr	rom			
M1000 fashion, and were matted together, indicating that they had been deposited by iver action, as opposed to having been deliberately   M1000 billog billog billog billog billog billog   M100 billog billog billog billog billog billog   billog Porland State University OR University billog billog billog   billog Geologiat Leonard Pairer PG Scientiat But the "free floating" hypothesis presented its own difficulties; it did not explain the ten bills missing from one packet, nor was there a logical reason that the the packets would have remained together after reparating from the rest of the money. Physical evidence was incomparatible with geologic evidence: finamelabach between the three-floating would have remained together after reparating from the rest of the money. Physical evidence was incomparatible with geologic evidence: finamelabach between the three-floating would have remained together after reparating from the rest of the money. Physical evidence was incomparatible with geologic evidence: finamelabach between the three-floating would have remained together after reparating from the rest of the money. Physical evidence was incomparatible with geologic evidence: finamelabach between the three-floating would have remained together after reparating from the rest of the money. Physical evidence was incomparatible with geologic evidence: finamelabach between the three-floating would have had to was have had	M:1004				one of its man	y connectin	g tributaries. An Army Corps of Engineers hydrologist noted that the bills	had disi	Integrated	i in a "ro	unded"				
M100 Columbia well downstream from the discovery site; and it lent credence to supplemental speculation (see later developments above) placing   M1006 b E103 Portland State University No   b E103 Geologist Leonard Palme PE Scientity   b E103 Geologist Leonard Palme PE Scientity   b E103 Himmebach PE Scientity But the "free packets owid have remained together after separating from the rest of the money. Physical evidence was   nonorpatible with the geologic evidence: Immelbach observed that there-floating hypotheres would have remained together after separating from the rest of the money. Physical evidence was	M:1009				fashion, and thuried [71] T	ere matted	together, indicating that they had been deposited by river action, as oppos firmed that Cooper had not landed near lake Merwin, nor in any other part of	ed to ha	wig Biver	delibera	tely eeds into	o the			
MUMD     Mumd     Constraint	M:1010				Columbia well	downstream	from the discovery site; and it lent credence to supplemental speculation (	see Late	er develop	ments abo	ve) place	ing			
PLIUS     Point and and comments     One     Owners     Description       PLIUS     Cologist Leand Palme     PER     Scientiat     a logical reason that the three packets would have remained together after separating from the rest of the money. Physical evidence was       b     Himmeldbach     PER     Scientiat     a logical reason that the three packets would have remained together after separating from the rest of the money. Physical evidence was	M:1016	Dealers & Grate Hairmaile	ORC	(Internation	the drop zone	near the Wa	shougal River, which merges with the Columbia upstream from the discovery s	ite.[72]							
bills Himmbach PR Scientitic a logical reason that the three packets would have remained together after separating from the rest of the money. Physical evidence size incompatible work of the gold reason that the three packets would have a for work on the set of the money. The back "within a couple the set of the money. The back "within a couple the set of th	D E-1019	Geologist Leonard Palmer	DER	Scientist	But the "free	floating" h	ypothesis presented its own difficulties; it did not explain the ten bills	missing	from one	packet, n	or was th	here			
incompatible with geologic evidence: Himmelsbach observed that free-floating bundles would have had to wash up on the bank "within a couple	D E:1018	Himmelsbach		Scientist	a logical reason that the three packets would have remained together after separating from the rest of the money. Physical evidence was										
PE1004 Portland, Oregon GPE of years" of the hijacking; otherwise the rubber bands would have long since deteriorated. [73] an observation confirmed experimentally by	▷ E:1004	Portland, Oregon			incomparate with geologic evidence: nimmelspace observed that ifeericating bundles would have had to Wash up on the bahk "within a could of years" of the hijacking; otherwise the rubber bahd would have long since deteriorated, (73) an observation confirmed experimentally the										
The El005 Seattle, Washington GPE the Cooper Research Team (see Ongoing investigation below). [74] Geologic evidence suggested, however, that the bills arrived at the area of	🗢 E:1005	Seattle, Washington	GPE		the Cooper Rea	search Team	(see Ongoing investigation below).[74] Geologic evidence suggested, however	, that t	he bills	arrived a	t the are	ea of			
Mi006 their discovery-a beach front known as Tina (or Tena) Bar-well after 1974, the year of a Corps of Engineers diredging operation on that article of the theory of Engineer of Portland State University found two distinct layers of and and engineer to between the clay	M:1006				their discove: stretch of the	y-a beach f. river, Geo	ront known as Tina (or Tena) Bar-well after 1974, the year of a Corps of En logist Leonard Palmer of Portland State University found two distinct layer	gineers	dredging	operation	on that	clay			
b E1006 FBI ORG deposited on the river bank by the dredge and the sand layer in which the bills were buried, indicating that the bills arrived long after	▷ E:1006	FBI	ORG		deposited on	he river ba	nk by the dredge and the sand layer in which the bills were buried, indicat	ing that	the bill	is arrived	long aft	ter			
P ELLOV American UPE dredging had been completed. [71][75] The Cooper Research Team later challenged Palmer's conclusion, citing evidence that the Casy layers	D E:1007	American GPE			dredging had been completed. [71] [75] The Cooper Research Team later challenged Palmer's conclusion, citing evidence that the clay layers over a neurosidence that the clay is the cooper Research Team later challenged Palmer's conclusion, citing evidence that the clay layers										
v timo 3 spress agent agent and the species into the species and the specie	+ E:1010	Portland International Airport	ORG		experiment), 1	out does not	help to explain how the bundles got to Tina Bar, or from where they came.[	76]	wased on	one rubbe	r Dang				
T Toblini	Tool-tine			-		-			Sea	rch		-			

Figure 2. This image shows a resident relation being added between mentions American (M:1008) and Cooper (M:1011)

Right-clicking on any mention brings up a tool bar that includes edit and delete options. The edit option allows users to change a value in the mention, so it does not need to be deleted if there is a typographical error. The most important information that is stored in this area is the relationships between mentions. Each mention acts as a parent node for the relationships that are associated with it. Each relationship acts as a child node, and holds the unique ID of the mention that shares the relationship.

The "Entities" tree is very similar in design to the "Mentions" tree, with a few key differences. The "Entities" tree stores unique entities from the text. Each entity has the same columns and options as mentions, but instead of holding relationships as child nodes, it holds unique mention identifiers for each occurrence of the mention. If a user clicks an entity, all of its mentions are highlighted in the "Mentions" tree, and if a user right-clicks, the edit and delete options appear just like the "Mentions" tree.

The text area displays the textual content of the loaded document. This is where the user selects text and either drag and drops it into the "Mentions" or "Entities" trees, or drags and drops it into a class button. This is how the user can create a relation between an entity and a non-entity string. A string can be selected in the text and added to a mention as a title or an age, to give two examples. The current document names are stored above the text. A button appears above the text area for each document that is loaded. These buttons can be used to quickly jump to different areas of the text.

The button bar provides easy access for most of the tasks this interface supports. There are some intuitive functions, such as New, Save, Load, Edit, Delete and Quit, as well as unique features like PER, ORG, GPE, Match and Relate. The PER, ORG, and GPE buttons are for creating new mentions with the chosen class type. If a user selects a name in text, then clicks one of these buttons, they will get a new mention with a unique ID, the name from text, the class from the button pressed, and a description that is manually entered. This also creates a new entity, unless the user chooses to add to existing entity. The match button will take all selected mentions and resolve them to the userdefined entity. Multiple mention selections can be made by holding down CTRL while clicking the mouse button. The relate button will make a relationship between two mentions. The user selects two mentions and presses the relate button. A prompt, as seen in Figure 2, will display asking which relationship to assign the mentions. There is also a help button that displays a list of frequently asked questions that, when clicked, display an answer. This help section can be seen in Figure 3.

The interface was created using Python 2.7.7. The modules used include Tkinter and ttk. Tkinter is a graphical user interface tool kit that includes some widgets such as buttons, list menus, and text areas. The module ttk, (Tkinter tool-kit) is an add on to the base Tkinter that provides some more advanced features and widgets. Python includes packages such as the Natural Language Tool-Kit (NLTK) that will be helpful in the automatic entity extraction process. Windows 7 was the operating system used for development, but testing was done in both Windows 7 and Ubuntu Linux 12.04.

The eventual goal for this system is to allow for automatic entity and relation extraction. It will also be a useful tool for semi-manual curation. This means users will be assisted by the natural language processing techniques while still extracting entities manually.

#### IV. USER STUDY

To evaluate the interactive tool, we carried out a formative user study that has aspects of a heuristic evaluation and aspects of a cognitive walkthrough [21]. In a heuristic evaluation [22], participants are typically experts in HCI and they rate how well the interface satisfies particular design principles, known as design *heuristics*. In a cognitive walkthrough [23], each participant is presented with a series of sub-tasks and they determine how easy or difficult it is to accomplish the task with the given interface. This is often done in the presence of the user study coordinator.

Because this was a formative user study rather than a large-scale task-based evaluation, participants were not given extensive tool training or demonstrations of its features. They were given a brief introductory text about the task, and presented with the interface, which includes help functionality. Most participants were experts in NLP and/or HCI. A defining characteristic of heuristic evaluations and cognitive walkthroughs is that they involve consultation with experts who may not be representative of an intended end-user.

Our user study was self-directed, with participants given several sub-tasks and asked questions about the ease or difficulty of accomplishing the task given the provided interface. All of the tasks concerned a short sample text excerpted from President Barack Obama's Wikipedia biography.

The tasks they were given were:

- Task 1. Load the provided text file.
- **Task 2.** Select all mentions of President Barack Obama (e.g. "President Obama," "Barack Obama", "Obama", etc.) and label each of them as PER (a person).
- **Task 3.** Ensure that all mentions of Barack Obama are matched to a single entity.
- **Task 4.** Select all mentions of Honolulu (e.g. "Honolulu" "Honolulu, HI", "Honolulu, Hawaii", etc.) and label each of them as GPE (geopolitical).
- **Task 5.** Ensure that all mentions of Honolulu are matched to a single entity.
- **Task 6.** For any portion of the text that describes Barack Obama as being born in Honolulu, add a *city-of-birth* relation between the Obama mention and the Honolulu mention in that portion of text.

76 ColdStart Gl	Л 1.2.1	-	and the local	ACCOUNT OF LAND	And Contractor	and the local division of the local division	-		-	spring integral 2 and integrations in A rank Party	Annual Street,	and so it is not	Second 1		j X
File															
	🖹 🗀 🏦										1			•	
New 9					Match	Relate	Relate					Delete	Delete		
		l.		C · /Ilsers		My School	1/NT.P. 10	ark-Study/DBCooperWikiTe	TDoc ty		de Marcu	<u>  .</u>			
Mentior	n ID's Name	Cla	Descriptio	t	, on 1 giro,	ing bonoc	27 1122 11	six coudy, proceptimizite.	JODGOI ON		eip menu				
M:1005	Portland, Oregon	GP								How can I select multiple items in the mentions/entity trees?	Select a	phrase in	the text,	then sel	then
M:1006	Seattle, Washingt	on GP							23	How do I add a new intertion.	press the	'Relate N	on-mentic	n' button	h. Or.
→ M:1007	FBI	OR		New	Doc1	Doc 2	Dor 3		Dump	How do I delete a mention?	select th	word or	phrase yo	u want to	add
org:mer	nbers M:10;	OR								How do I delete an entity?	to a ment	ion. Then	right cli	ck and dr	cag the
M:1008	American	GP		analysis	agreed t	that it w	ould mal	ke little difference, [63	as 🔺	How do I turn on/off tooltips?	selection	to the de	sired men	tion. Fol	llow
M:1009	Boeing 727 aircra	ft OR		interpre	station of	f the sta	tute va:	ries considerably from c	ase to	How do I add a relation between two mentions:	the promp	ta to comp	Tere rue	cask.	
M:1010	aircraft	PEF		case and	d court to	court,	and a pr	rosecutor could argue the	at	How can I add a non-mention relation (i.e pertitle)?					
M:1011				Cooper 1	had forfei	ited immu	nity on	any of several valid							
M:1012	the agency	OR		November	r when a l	Portland	grand to	irv returned an indictme							
▷ M:1013	Special Agent Lar	ry Ca PEF		against	"John Doe	, aka Da	n Cooper	r", in absentia for air p	piracy -						
M:1014	Portland Internat	onal OR		and viol	lation of	the Hobb	a Act. [	66] The indictment in ef	fect						
▷ M:1015	Dan Cooper	PEF		formally	/ initiate	ed prosec	ution of	f the hijacker that can l	be						
▷ M:1016	Flight 305	OR		Continue	ed, should	i he be a	pprehend	ied, at any time in the							
Faile IDia		Ch.		rucure.	[04]										
Entity ID's	Name	Cia	rescription	Physical	l evidence										
₩ 480 E:1001	D. B. Cooper	PEP		In 1978	a placard	i contain	ing inst	tructions for lowering th	he aft						
M:1001				about 13	Smiles (3	as round	at of C	er nunter near a loggin	g road						
M:1011		10255		north of	f Lake Mer	rwin, but	within	the basic path of Flight	C C						
D E:1002	hijacked	PEF		305.[67]											
⇒ E:1003	Boeing	OR													
M:1004				Portion	of Brian	Ingramia	1090 4	A COMPANY							
M:1009				In Febru	ary 1980	an eight	-vear-o	ld boy named Brian Ingra	η.						
M:1010				vacation	ning with	his fami	ly on th	ne Columbia River about	9						
M:1016				miles (1	14 km) dos	mstream	from Van	ncouver, Washington, and	20						
▷ E:1020	Portland State University	OR U	niversity	miles (S	32 km) sou	athwest c	f Ariel,	uncovered three packet.	s of						
▷ E:1019	Geologist Leonard Palmer	PEF S	tientist	in rubbe	som cash,	signific	antly di	sandy riverbank to built	anaiea						
▷ E:1018	Himmelsbach	PEF S	ientist	campfire	. [68] FBI	I technic	ians con	firmed that the money w	15						
▷ E:1004	Portland, Oregon	GPI		indeed a	a portion	of the r	ansom-ti	wo packets of 100							
₩ E:1005	Seattle, Washington	GPI		twenty-c	iollar bil	lls each,	and a t	third packet of 90, all							
M:1006				arranged	i in the s	same orde	r as whe	an given to Cooper.[69][	70]						
▷ E:1006	FBI	OR		The disc	overy la	inched se	veral n	w rounds of conjecture.	and						
▷ E:1007	American	GPI		ultimate	aly raised	i many mc	re quest	tions than it answered.							
▷ E:1009	Special Agent Larry Carr	PEF		Initial	statement	ts by inv	estigat	ors and scientific consul	ltants						
	Portland International Airpor	t OR		were for	inded on t	the assum	ption th	hat the bundled bills was	shed						
Test for				Treelu 1	INTO THE (	o umbia	WIVER P	com one of its many conn	Porting .			<b>C</b> -1		_	-
rool-tips						_						26	iren:		

Figure 3. This image shows the help area. Several questions are displayed with their answers.

• **Task 7.** Locate a portion of the text that describes Barack Obama as having the title President of the United States. Add a *title* relation between the mention of Obama and the string "President of the United States."

#### • Task 8. Save the annotations.

For each of the above tasks, the participants were asked to answer the following questions:

- Is the task clear?
- Does the interface provide appropriate tools for accomplishing the task easily?
- Is the interface missing any tools for accomplishing this task?
  - If so, what?
- Is the system feedback appropriate? For example, were you confident that you had successfully finished the task?
- Do you have any suggestions for making this task easier?

Additionally, they were asked to provide several Likertscale ratings:

- How easy to use is the user interface, on a scale of 1-7?
- How attractive is the user interface, on a scale of 1-7?
- How efficient are the tasks overall using this interface, on a scale of 1-7?

On the 1-7 scale for each of the criteria above, 1 corresponds to *not at all* while 7 is *very*.

We also asked our participants how much experience they have in NLP and in HCI. Finally, we asked if they had any further open-ended comments or suggestions.

The user study was carried out in two locations: Sydney, Australia and Abbotsford, BC, Canada. Educational levels ranged from first-year undergraduate to doctoral students. Not all participants were native English speakers, but all were very proficient in English.

## V. RESULTS

Figure 4 summarizes results. The tasks varied on how easy they were to complete. All participants completed trivial tasks like loading and saving easily, but other tasks were varied in their ease. For example, task 7, described in Section IV and requiring the participant to relate the Barack Obama entity to a string representing his title, was not completed easily by most participants. The challenge is that this is a relationship between an entity and a string rather than a relationship between two entities, and most participants asked for "a more explicit procedure", or that the interface prompts be provided "more intuitively." This shows that better feedback is needed and more steering as far as leading users in the correct direction. One participant gave a suggestion about the provided help area: "improve the help menu with a step-by-step description of how to do the tasks." This shows that users require a more detailed explanation of the task, not just which button they need to press.

Other tasks were completed easily by some and not by others, including task 3, which asked users to "Ensure that



Figure 4. Likert-Scale Ratings by Participant



Figure 5. Ease of Use vs knowledge of natural language processing

all mentions of Barack Obama are matched to a single entity." Some participants could quickly see that the mentions were properly matched to their corresponding entity, and that when they clicked on the entity, a link icon is displayed beside all matched mentions. Other participants could not complete this task, saying simply "I got lost. If there was an easier pathway to do it, that would be great." This is proof that even with the help button present on the main surface of the interface, not all users will search the help area for information about their task. In the case of task 3, there was a help section labeled "How do I match a mention to an entity?"

Another finding from this study is that a good grasp of natural language processing and the KBP curation task is very helpful when using the system. This can be seen in Figure 5. Many of the participant comments dealt with the terminology and structure of the tasks and what the task requires. One participant asked the question, "should the President of the United States string be added as an Organization entity before adding the title relation?" If the user had known that an attribute like *age* or *title* is not a relationship between two named entities, then they would have found the "relate string" button. The purpose of this interactive tool is not to teach users about the terminology or the curation of KBP, but instead, it is designed to aid users who are already doing this task.

Most participants wanted more specific system feedback and success prompts, such as "Mention matched successfully to entity" or "New relation created." This would give users a better sense that they are carrying out the tasks correctly. More specifically, some users were not happy with the number of pop-ups they were interacting with. The challenge here is to create better feedback, while reducing the number of prompts. One solution that was suggested by participants in the study was the use of drag-and-drop to match mentions and entities and also to relate mentions. Drag and drop reduces the number of times the interface needs to ask for mention or entity choices, and also speeds up the process of matching or relating mentions.

A strong point of the interactive tool is that regardless of the issues that the participants were having, most of them were still able to produce a correct mention and entity tree. This is evident from the saved files that some of the participants made, and the comments from the others. This suggests that the system is robust enough to handle user experimentation and still prevent enough errors that the user eventually will find the correct method without losing work they have already done. The only task that some participants eventually gave up on was task 6, where it asks users to add a *city-of-birth* relation between the Obama mention and the Honolulu mention. Some users never found that to select more than one mention, the user has to hold the CTRL button. This led to comments like, "I tried selecting 2 entities to relate them but it only let me select 1."

As is evident in Figure 5, the higher the participant's knowledge of natural language processing, the easier the tasks were to complete.

## VI. FUTURE WORK

The results of the user study suggest a number of refinements for the document annotation interface.

Assisted curation: We are currently adapting automatic methods for entity and relationship annotation. We will incorporate these into the user interface as suggestions for human correction, turning our interactive tool into an intelligent user interface (IUI). Rather than annotating everything manually, users will correct output from the machine. This raises new questions about presentation and prioritisation. How should we compute (lack-of) confidence from the machine and convey that to users in the interface? Can machine confidence be used to confidently add some annotations automatically with acceptable precision? Can machine confidence be used to confidently ignore some annotations with acceptable recall? What is the best way to prioritise the remaining machine annotations for human review?

*KB-centric curation for cold-start KBP:* We will extend the user interface from a document-centric tool for entity linking to a KB-centric tool for building a KB from scratch. This requires new functionality, e.g., processing and browsing multiple documents, prioritising curation tasks across the document collection, presentation and navigation of curation tasks, functionality to apply curation actions across mentions and documents, and KB-update and tool update mechanisms. We will evaluate effectiveness using the cold-start entity discovery and slot filling tasks from this year's Text Analysis Conference.<sup>1</sup> Specifically, we will explore the tradeoff between cost, coverage, and precision. We will also benchmark interactive KBP against state-of-the-art methods for automatic KBP.

*Task-Based Evaluation:* The formative user study has provided guidance for subsequent large-scale task-based evaluations we will carry out. In those evaluations, we will ensure participants are given training time and demonstration of the tool features. They will also be thoroughly versed in the knowledge base population task. In those studies, we will aim to capture more detailed information such as the time taken to complete a task, and the number of errors made while completing a task. We also plan to instrument the interface in order to capture and log information about where the participants are clicking.

*Querying:* In addition to facilitating the creation of a knowledge base from unstructured text, we want to allow the user to query the knowledge base in order to satisfy an information need. To continue our running example, a sample query may be "Who has the title President of the United States?" Some queries may require the IUI to reason about several relations in combination with one another.

Finally, we will make the interface software and source code available upon publication of this paper.

## VII. CONCLUSION

We have described the development of an interactive tool to assist manual creation of a knowledge base from an input consisting of unstructured text. The tool can be used for identifying mentions of entities, labeling their entity type, matching mentions that refer to the same entity, and adding relations or facts about entities. This is a valuable and needed resource for a domain where fully-automated systems still struggle with low accuracy.

The formative user study results were encouraging in that all participants were able to create and save a knowledge base from a sample text input. The participants who were most familiar with NLP found the tool to be appropriate and efficient for the task, indicating that this will be an effective annotation tool for NLP researchers.

Outside of the NLP research community, we foresee this software being useful in business intelligence and enterprise data management. Many organizations possess unstructured data that is not fully exploited. Being able to build a knowledge base from that data, and having the functionality to query that knowledge base, would allow organizations to incorporate the data into their decision-making processes.

#### ACKNOWLEDGMENT

Ben Hachey is the recipient of an Australian Research Council Discovery Early Career Researcher Award (DE120102900).

<sup>1</sup>http://www.nist.gov/tac/2015/KBP/ColdStart/

#### REFERENCES

- M. Surdeanu and H. Ji, "Overview of the English slot filling track at the TAC2014 knowledge base population evalution," in *Proceedings of the Text Analysis Conference*, 2014.
- [2] N. Karamanis, I. Lewin, R. Seal, R. Drysdale, and E. Briscoe, "Integrating natural language processing with Flybase curation," in *Proceedings of the Pacific Symposium on Biocomputing 12*, 2007, pp. 245–256.
- [3] B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, and X. Wang, "Assisted curation: does text mining really help?" in *Proceedings of the Pacific Symposium on Biocomputing* 13, 2008, pp. 556–567.
- [4] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: a web-based tool for NLP-assisted text annotation," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 102–107.
- [5] H. Ji and R. Grishman, "Knowledge base population: Successful approaches and challenges," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 1148–1158.
- [6] H. Ji, J. Nothman, and B. Hachey, "Overview of TAC-KBP2014 entity discovery and linking," in *Proceedings of* the Text Analysis Conference, 2014.
- [7] J. Guo, G. Xu, X. Cheng, and H. Li, "Named entity recognition in query," in *Proceedings of the 32nd International Conference on Research and Development in Information Retrieval*, 2009, pp. 267–274.
- [8] G. Angeli, J. Tibshirani, and J. Y. Wu, "Combining distant and partial supervision for relation extraction," in *Proceedings* of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1556–1567.
- [9] Z. Lu and L. Hirschmann, "Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II," *Database*, 2012, accessed 15 June 2015 from http://database.oxfordjournals.org/content/2012/bas043.full.
- [10] D. McDonald and U. Kelly, "Value and benefits of text mining," Jisc, Tech. Rep., 2012, accessed 15 June 2015 from http: //www.jisc.ac.uk/reports/value-and-benefits-of-text-mining.
- [11] L. Hirschman, G. A. P. C. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, A. Lourencço, R. Nash, A.-L. Veuthey, T. Wiegers, and A. G. Winter, "Text mining for the biocuration workflow," *Database*, 2012, accessed 16 June 2015 from http://database.oxfordjournals.org/content/ 2012/bas020.full.
- [12] C.-H. Wei, B. R. Harris, D. Li, T. Z. Berardini, E. Huala, H.-Y. Kao, and Z. Lu, "Accelerating literature curation with textmining tools: a case study of using PubTator to curate genes in PubMed abstracts," *Database*, 2012, accessed 16 June 2015 from http://database.oxfordjournals.org/content/2012/bas041. full.

- [13] C. N. Arighi, B. Carterette, K. B. Cohen, M. Krallinger, W. J. Wilbur, P. Fey, R. Dodson, L. Cooper, C. E. V. Slyke, W. Dahdul, P. Mabee, D. Li, B. Harris, M. Gillespie, S. Jimenez, P. Roberts, L. Matthews, K. Becker, H. Drabkin, S. Bello, L. Licata, A. Chatr-aryamontri, M. L. Schaeffer, J. Park, M. Haendel, K. V. Auken, Y. Li, J. Chan, H.-M. Muller, H. Cui, J. P. Balhoff, J. C.-Y. Wu, Z. Lu, C.-H. Wei, C. O. Tudor, K. Raja, S. Subramani, J. Natarajan, J. M. Cejuela, P. Dubey, and C. Wu, "An overview of the BioCreative 2012 Workshop Track III: interactive text mining task," *Database*, 2012, accessed 16 June 2015 from http://database.oxfordjournals.org/content/2013/bas056.long.
- [14] F. Niu, C. Zhang, C. Ré, and J. Shavlik, "Elementary: largescale knowledge-base construction via machine learning and statistical inference," *International Journal on Semantic Web & Information Systems*, vol. 8, no. 3, pp. 42–73, 2012.
- [15] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell, "Toward an architecture for neverending language learning," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010, pp. 1306–1313.
- [16] N. Nakashole, M. Theobald, and G. Weikum, "Scalable knowledge harvesting with high precision and high recall," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, 2011, pp. 227–236.
- [17] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: a web-scale approach to probabilistic knowledge fusion," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 601–610.
- [18] K. Bontcheva, I. Roberts, L. Dercynski, and D. Rout, "The GATE crowdsourcing plugin: crowdsourcing annotated corpora made easy," in *Proceedings of the Demonstrations at the* 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 97–100.
- [19] S. M. Yimam, I. Gurevych, R. E. de Castilho, and C. Biemann, "WebAnno: A flexible, web-based and visually supported system for distributed annotations," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2013, pp. 1–6.
- [20] W. Radford, D. Tse, J. Nothman, B. Hachey, G. Wright, W. C. James R. Curran, T. O'Keefe, M. Honnibal, D. Vadas, and C. Loxley, "The Computable News project: research in the newsroom," in *Proceedings of the WWW Workshop on Web and Data Science for News Publishing*, 2015, pp. 903–908.
- [21] Y. Rogers, H. Sharp, and J. Preece, *Interaction Design*, 3rd ed. West Sussex, UK: Wiley, 2011.
- [22] J. Nielsen, "Heuristic evaluation," Usability inspection methods, vol. 17, no. 1, pp. 25–62, 1994.
- [23] —, "Usability inspection methods," in *Conference compan*ion on Human factors in computing systems. ACM, 1994, pp. 413–414.