

**ENTITY-BASED ENRICHMENT FOR INFORMATION
EXTRACTION AND RETRIEVAL**

A Dissertation Presented

by

JEFFREY DALTON

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2014

Computer Science

© Copyright by Jeffrey Dalton 2014

All Rights Reserved

**ENTITY-BASED ENRICHMENT FOR INFORMATION
EXTRACTION AND RETRIEVAL**

A Dissertation Presented

by

JEFFREY DALTON

Approved as to style and content by:

James Allan, Chair

David Smith, Member

W. Bruce Croft, Member

Rajesh Bhatt, Member

Laura Dietz, Member

Lori A. Clarke, Department Chair
Computer Science

*Dedicated to my wife, Krystle
And my parents, Jon and Karen.*

ACKNOWLEDGMENTS

There are many people that I would like to thank for helping throughout this process. Without their help this thesis would not have been possible.

First and foremost, I would like to thank my advisor, James Allan. I would like to thank James for his assistance and guidance. James taught me important lessons about the scientific process and research. His repeated admonitions to look at the data and dig deeper shaped my research. I would also like to thank him for his flexibility and patience, including a brief foray into reality television. James' humor and steadfastness made this thesis possible.

I'd also like to thank Laura Dietz for being a partner in research and close friend. I want to thank her countless hours of discussion, debate, writing, and coding, that were instrumental in this work. I also want to thank Laura for her enthusiasm and encouragement.

I would also like to thank David A. Smith for his guidance and inspiration. His depth of knowledge across diverse fields was invaluable and strongly influenced the direction of my research. I would also like to thank Bruce Croft for his insightful questions and deep knowledge of prior work that improved this work. I would like to thank Andrew McCallum for his help and advice. I would also like to thank my committee member Rajesh Bhatt for his insightful questions and assistance.

I'd like to thank the graduate students and CIIR alumni who made the research lab an inspiring and unforgettable experience. I learned a lot from our many discussions, debates, and arguments about information retrieval. I would like to thank them for taking the time to listen to practice talks, refine ideas, and help with experiments. These include: Elif Aktolga, Niranjan Balasubramanian, Michael Bendersky, Ethem Can, Marc Cartright, Van Dang,

Shiri Dori-Hacohen, Henry Feild, Logan Giorda, John Foley, Sam Huston, Myung-ha Jang, Jin Young Kim, Kriste Krstovski, Mostafa Keikha, Weize Kong, Chia-Jung Lee, Matt Lease, Tamsin Maxwell, Jae Hyun Park, Jangwon Seo, David Wemhoener, Xiaobing Xue, Xing Yi, and everyone else. I also want to thank the graduate students in IESL that I had the pleasure to work with: Sam Anzaroot, Anton Bakalov, David Belanger, Daniel Duckworth, Ariel Kobren, Alexandre Passos, Sameer Singh, Michael Wick, and Limin Yao.

I had the opportunity to spend the summer at several internships. My summer at Yahoo! Research in Barcelona with the NLR group shaped my time. I'd like to thank Peter Mika and Roi Blanco for their guidance, advice, and foosball matches. I'd like to thank Vanessa Murdock for her advice and friendship.

My internship with the search quality at Twitter was an unforgettable experience. I'd like to thank my mentors there, Gilad Mishne and Aneesh Sharma. They provided me with valuable insights on aspects of applying research to the real-world and the satisfaction of shipping features used by millions of people. I'd also like to thank my other colleagues: Abdur Chowdhury, Jimmy Lin, Michael Paul, Stanislov Nikolov, Andy Schlaikjer, and Zhengua Li.

Prior to joining CIIR, I was fortunate enough to have the experience building search engines at Globalspec. I would like to thank Steinar Flatland and Steve MacMinn, who gave me the opportunity to be part of the search engine development team. They introduced me to search and supported my decision to pursue research. I would also like to thank my other colleagues and friends who supported me, including: Joan, Jeremy, Rich, Dan, and Kathleen.

I'm sincerely thankful to all the staff who have helped me during my time. In particular, I'd like to thank Kate Moruzzi for her eager willingness to help with so many day-to-day tasks. I'd like to thank Dan Parker and Andre Gauthier for their technical help and support with our clusters and computers. I'd like to thank Jean Joyce and Glen Stowell for their help. I would like to thank David Fisher for his thoughts and invaluable depths of knowledge. I

would like to thank Barbara Sutherland for welcoming me many mornings with a smile, which brightened difficult times. A special thanks to Leeanne Leclerc for her tireless efforts keeping everything organized and for her help navigating the maze of university and department requirements.

I'd like to thank all of the friends I've made through this process and for making this place an enjoyable place to live. These include: James and Kristen Atwood, Ben Bayes and Heather Lockrow, John Bowers, Marco Carmosino, Will Dabney, Jacqueline Feild, Luisa Galindo, Dubi Katz, Marina Levin, Bridget MacDonald, Brandon McPhail, and Mike and Kristin Thomas. I'd like to thank everyone at MERCYhouse whose thoughts and prayers have been a sustaining force during my time here, including: Felicia Bokel, Christopher Boulton, Tracy Conner, Dan Ratelle, Bill Coolley, Lois Grandmaison, and Brett and Jenna Marquard.

I am very thankful for the support of my family, my parents Jon and Karen. They inspired me to pursue learning. I would like to thank them for their unconditional love and support. Most of all, I owe an immeasurable amount of thanks to my wife and best friend, Krystle, who supported me from the beginning. Without her love and partnership this thesis would not be possible. Thank you all for making this possible.

This work was supported in part by the Center for Intelligent Information Retrieval, in part by IBM subcontract #4913003298 under DARPA prime contract #HR001-12-C-0015, in part under subcontract #19-000208 from SRI International, prime contractor to DARPA contract #HR001-12-C-0016, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, in part by NSF CLUE IIS-0844226, in part by a grant from the Andrew W. Mellon Foundation, and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect those of the sponsor.

ABSTRACT

ENTITY-BASED ENRICHMENT FOR INFORMATION EXTRACTION AND RETRIEVAL

MAY 2014

JEFFREY DALTON

B.Sc., UNION COLLEGE

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor James Allan

The goal of this work is to leverage cross-document entity relationships for improved understanding of queries and documents. We define an entity to be a thing or concept that exists in the world, such as a politician, a battle, a film, or a color. Entity-based enrichment (EBE) is a new expansion model for both queries and documents using features from similar entity mentions in the document collection and external knowledge resources. It uses task-specific features from entities beyond words that include: name aliases, fine-grained entity types, categories, and relationships to other entities. EBE addresses the problem of sparse or noisy local evidence due to multiple topics, implicit context, or informal writing.

With the ultimate goal of improving information retrieval effectiveness, we start from unstructured text and through information extraction build up rich entity-based representations linked to external knowledge resources. We study the application of entity-based enrichment to each step in the pipeline: 1) Named entity recognition, 2) Entity linking, and

3) Ad hoc document retrieval. The empirical results for EBE in each of these tasks shows significant improvements.

Our first application of entity-based enrichment is the task of detecting entities in named entity recognition. We enrich the representation of observed words likely to represent entities. We perform weighted feature copying of recognition features from similar tokens in the corpus and external collections. The evaluation shows statistically significant improvements on in-domain newswire accuracy and demonstrates that the models are more robust on out-of-domain data.

In the second part of this work, we apply EBE to the task of entity linking. The proposed entity linking method for disambiguating the detected mentions to entries in an external knowledge base is based on information retrieval. The neighborhood relevance model, an enrichment model, identifies salient associations between an entity mention and other entity mentions in the document. The results show that the enrichment model outperforms other context models and results in a system that is competitive with leading methods.

Using the constructed entity representation, the final task is ad hoc document retrieval. We enrich the query representation with entity features. Retrieval is performed over documents annotated with entities linked to Wikipedia and Freebase from our system as well as the publicly available Google FACC1 annotations. To effectively leverage linked entity features, we extend dependency-based retrieval models to include structured attributes. We also define a new query-specific entity context model which builds a model for disambiguated entities from retrieved documents. Our results demonstrate significant improvements over competitive query expansion baselines for several standard test collections.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	v
ABSTRACT	viii
LIST OF TABLES	xv
LIST OF FIGURES	xviii
 CHAPTER	
1. INTRODUCTION	1
1.1 Entities in Retrieval	2
1.2 Beyond Entity Retrieval	4
1.3 Opportunity of Entity Representations	6
1.4 Text to Entity Representations	8
1.5 Entity-based Enrichment	9
1.6 Tasks	10
1.6.1 Named Entity Recognition	10
1.6.2 Entity linking	11
1.6.3 Document Retrieval	11
1.7 Contributions	12
1.7.1 Outline	14
2. BACKGROUND	15
2.1 Query Expansion	15
2.2 Document Expansion	18
2.3 Word Sense Disambiguation	19
2.4 Controlled Vocabulary Representations	20
2.4.1 Structured Classification	20

2.4.2	Wikipedia	22
2.4.3	Implicit Topics	23
2.5	Entities in Retrieval	23
2.6	Graphical models for IR	25
2.6.1	Log-linear Models	26
2.6.2	Retrieval Models	26
2.6.3	Relevance Feedback	27
3.	DATA	29
3.1	Named Entity Recognition Corpora	29
3.1.1	CoNLL 2003	29
3.1.2	Deerfield Book Collection	30
3.2	Entity Linking	30
3.3	Ad Hoc Retrieval	33
3.3.1	Documents	33
3.3.2	Topics	34
3.3.3	Relevance Assessments	34
3.3.4	Entity Annotations	34
3.4	Evaluation	36
3.4.1	Named Entity Recognition	37
3.4.2	Entity Linking	37
3.4.3	Retrieval	38
4.	ENTITY-BASED ENRICHMENT	42
4.1	Terminology	42
4.2	Overview	43
4.3	Enrichment Triggering	44
4.4	Target Model Generation	45
4.5	Mention Retrieval	46
4.6	Mention Feature Extraction	47
4.7	Feature Aggregation	48
4.8	Summary	49
5.	NAMED ENTITY RECOGNITION	52
5.1	Introduction	52
5.2	Non-Local Dependencies in NER	55

5.3	NER Approach	57
5.4	Entity-based Enrichment for NER.....	59
5.4.1	Enrichment Scope	59
5.4.2	Enrichment Triggering	60
5.4.3	Target Model Generation	61
5.4.4	Mention Source Retrieval	62
5.4.4.1	Exact match	62
5.4.4.2	Unigram	63
5.4.4.3	Term Dependence Models.....	63
5.4.5	Mention Feature Extraction	63
5.4.6	Feature Aggregation	64
5.5	Experiments	64
5.5.1	Enrichment Triggering Evaluation	65
5.5.2	Source Retrieval Effectiveness	66
5.5.3	CoNLL NER Evaluation.....	68
5.5.3.1	Local NER Models.....	68
5.5.3.2	Exact Match Feature Enrichment	70
5.5.3.3	Ranked Feature Enrichment	72
5.5.4	Deerfield Evaluation	73
5.5.4.1	Exact Match Enrichment	73
5.5.4.2	Ranked Feature Enrichment	74
5.5.5	Enrichment using External Collections	74
5.5.5.1	Reuters RCV1 subset	74
5.5.5.2	Evaluation	75
5.6	Summary	75
6.	ENTITY LINKING	77
6.1	Introduction	78
6.2	Related Work	81
6.3	Mention Context Model	82
6.3.1	Local Neighborhood Model	84
6.4	Entity Linking Retrieval Model	85
6.5	Entity-based Enrichment for Linking	87

6.6	KB Bridge: Entity Linking System	89
6.6.1	Knowledge Base Representation	90
6.6.2	Document Analysis	90
6.6.3	KB Entity Ranking	91
6.6.4	NIL Handling	92
6.7	Experimental Evaluation	92
6.7.1	Target Model Evaluation	92
6.7.2	Ranking Evaluation	93
6.7.2.1	Recall	95
6.7.3	TAC KBP results	96
6.8	Enrichment approaches for TAC KBP 2013	97
6.8.1	Urban Dictionary Enrichment	98
6.8.2	Entity KB Coherence	98
6.8.3	Results	100
6.9	Summary	101
7.	ENTITY-BASED FEATURE ENRICHMENT FOR RETRIEVAL	102
7.1	Entity-Based Document Retrieval	105
7.1.1	Cross-vocabulary dependencies	107
7.2	Entity Context Model	109
7.2.1	Related entity models	111
7.3	Entity Query Feature Enrichment	112
7.4	Experimental Setup	114
7.5	Experimental Evaluation	115
7.5.1	Overall Performance of EQFE	115
7.5.2	Entity Analysis of queries	119
7.5.3	Effectiveness by type on Robust04	120
7.5.4	Feature-by-Feature Study	121
7.5.5	Error Analysis of ClueWeb09	122
7.6	Summary	126

8. CONCLUSION	128
8.1 Future Work	130
BIBLIOGRAPHY	135

LIST OF TABLES

Table	Page
3.1	NER collection statistics 29
3.2	TAC Source Corpus 31
3.3	TAC Knowledge Base Types 31
3.4	TAC Mention Types by Year 31
3.5	Test Collections Statistics 35
3.6	FACC1 ClueWeb Annotation Statistics 36
5.1	Baseline NER features 58
5.2	Query trigger features 59
5.3	Query Trigger evaluation on the CoNLL training data. It compares boolean combinations of the features from Table 5.2. 65
5.4	Evaluation of case normalization in retrieval using the Query Likelihood ranking and no context for the 33,429 queries. 66
5.5	Evaluation of sentence retrieval using Mean Average Precision (in %). Various combinations of case sensitivity, retrieval model, and query generation method are evaluated. QL indicates Query Likelihood retrieval, SD indicates Sequential Dependence. The last word indicates the query generation method from Section 5.4.3. 66
5.6	Phrase level F1 scores for base NER models described in Section 5.3 compared with the Stanford NER tagger on the CoNLL 2003 Named Entity Recognition test (b) set. 67

5.7	F1 scores on CoNLL for feature enrichment using exact string matching for varying corpus scopes described in Section 5.4.3. The top is the baseline model with features from Table 5.1. The bottom results are for a stronger model with Brown clusters and Wikipedia features. Statistically significant over local models where indicated with a * with $p \leq .05$	69
5.8	CoNLL F1 scores for feature enrichment using ranked source retrieval with the Global retrieval scope. (QL) indicates Query Likelihood and (SD) indicates Sequential Dependence retrieval models. The query context models used two variations: Capitalized includes only capitalized tokens, All has all tokens excluding stopwords. Significant differences over the local model with $p \leq .05$ are indicated with by *.....	69
5.9	F1 scores of the NER model trained on CoNLL and evaluated on the Deerfield collection. The results show local systems and unweighted feature enrichment with varying collection scopes. The top is a tagger model with baseline features. The bottom is a stronger baseline model with word clustering and Wikipedia features. The differences are statistically significant with local models where indicated with a * with $p \leq .05$	70
5.10	F1 scores for CoNLL models evaluated on the Deerfield collection. The table compares global ranked feature enrichment models compared with baseline exact string matching. We compare against the state-of-the-art LBJ NER model that uses Fixed Window feature aggregation. All differences are statistically significant over the baseline ExactMatch model with a with $p \leq .05$, a * indicates significance over LBJ.	71
5.11	F1 scores for external feature enrichment including a 50k document subset of the RCV1 reuters news collection. Ext100 indicates 100 feedback sentences, Ext50 indicates 50 sentences. A * indicates significance over non-external model with $p \leq .05$	72
5.12	Summary Table comparing the F1 score of the strongest models in each category, a purely local model incorporating word clustering and gazetteers, a model using ranked feature enrichment models, and enrichment including an external corpus. All results are statistically significant with $p \leq .05$	72
6.1	Features of the query mention and candidate Wikipedia entity.	91

6.2	Ranking results on TAC by year with varying context methods with mean reciprocal rank (MRR). The best results for each year are highlighted in bold. Results that are statistically significant with $\alpha = 5\%$ over the QV baseline are indicated with *.	95
6.3	Learning to rank refinement results with mean reciprocal rank (MRR). All LTR results are statistically significant with $\alpha = 5\%$ over the unsupervised QVM_nrm	95
6.4	TAC Entity Linking performance in macro-average accuracy.	96
6.5	Features of the entity-to-entity similarity.	99
6.6	Overall effectiveness in 2013.	100
6.7	B ³⁺ F1 by document type.	100
6.8	B ³⁺ F1 by entity class.	100
7.1	Example expansion terms for the query “Obama Family Tree”	115
7.2	Summary of results comparing EQFE with other methods across the three test collections.	117
7.3	Queries EFQE helped versus hurt over SDM baseline.	117

LIST OF FIGURES

Figure	Page
3.1 Example Robust Query	36
3.2 Example FACC1 Entity Annotations	36
4.1 Passage from Old Paths of the New England Border, pages 164,166.	42
4.2 Glossary of terms	43
4.3 Example context model for the observed token ‘Hadley’ in entity recognition.	45
4.4 Example mention retrieval query constructed from the ‘Hadley’ context model	46
4.5 Example retrieved passages for the ‘Hadley’ target.	50
4.6 Subset of extracted entity recognition features f_m from related mention. 51	
4.7 Sample aggregated feature values recognition features f_{En}	51
6.1 Excerpt from TAC document with linking query for [Toon] entity.	77
6.2 Example context model for [Toon] entity target from Figure 6.1	84
6.3 Contextual query for [Toon], occurring in the sentence from Figure 6.1	87
6.4 Ablation study for the suggested method in terms of Precision @ 1.	94
6.5 Average recall at rank cutoff k.	96
7.1 Example Freebase entity for Barack Obama, /m/02mjmr	106
7.2 FACC1 entity document annotations for clueweb09-en0004-08-20390	107

7.3	FACC1 entity annotations for TREC Web Track query 1: [obama family tree]	107
7.4	Overview over feature sources.	108
7.5	Example expansion of query C09-1 with entities [] and Freebase types {}.....	109
7.6	Mean retrieval effectiveness with standard error bars.....	116
7.7	Mean retrieval effectiveness across different query-difficulties, measured according to the percentile of the SDM method. (The hardest queries are on the left)	118
7.8	Number of queries containing different classes of entities (manual labeling) 119	
7.9	Percentage of queries containing different classes of entities	119
7.10	Mean Average Precision over different classes of entity queries on Robust04	120
7.11	Features sorted by retrieval effectiveness on Robust04.	123
7.12	Features sorted by retrieval effectiveness on ClueWeb09B.	124
7.13	Features sorted by retrieval effectiveness on ClueWeb12B.	125

CHAPTER 1

INTRODUCTION

Word-based representation of documents in the field of information retrieval have proven to be effective over diverse collections. In their simplest form, documents are modeled as a bag-of-words, where each term is independent of other terms in the document. Beyond bag-of-words models, recent efforts have focused on leveraging dependencies between terms (METZLER and CROFT 2005) to model phrases and proximity. However, as search applications evolve and become more complex, representing documents solely using words is limited. Words alone do not support joins or provide structured relationships needed for more complex inference. Entity-based representations combine both text and structured entity attributes. The quantity and simplicity of text documents combined with entity annotations supports more complex queries and cross-document inference.

One of the primary motivations of this thesis is to leverage newly available entity knowledge bases to enhance the representation of documents and queries. The use of entities in retrieval has a long history, which we describe in more detail in Chapter 2. But, the availability of large-scale general purpose knowledge bases is a recent development. The knowledge bases we focus on in this thesis, as well as others, gained prominence in the late 2000s. Many build on the success of Wikipedia and further expand the information from it. These differ from previous knowledge bases because they focus on coverage of both general concepts as well as named entities. The research in this thesis is enabled by these resources. We expect the development of these and similar knowledge bases to evolve and become more important, especially with recent efforts on improving automated methods for knowledge base construction (SUCHANEK *et al.* 2013).

The focus of this dissertation is understanding relationships across documents using entities. We define an *entity* broadly to be a thing or concept that exists in the world, such as a person, a battle, a film, or a color. Entities exist as mentions across documents and in external knowledge resources. The goal of this work is to construct and utilize entity-based representations to improve effectiveness for a variety of document analysis tasks in information extraction and information retrieval.

The challenge we address in this thesis is how to augment text using entity representations for a particular task, a process we refer to as *entity enrichment*. Entities are the basis for enrichment because they are a unit of meaning shared across documents and in external knowledge resources. Entity enrichment is a process that includes text expansion (SINGHAL and PEREIRA 1999; TAO *et al.* 2006a), but focuses on structured feature expansion of local observations from similar entity mentions in other documents and in external knowledge base entries.

1.1 Entities in Retrieval

The recent popularity of question answering services such as IBM Watson and personal assistants including Siri and Google Now are part of an increasing trend of search engines returning answers to users. Answers are often an entity, an entity attribute, or a list of entities. Entity results are retrieved from a wide variety of structured databases. Google incorporates entity data from their Knowledge Graph and Google Plus. Yahoo! uses the Web Of Objects. Bing returns Facebook and Satori entities. Facebook Graph Search performs retrieval over its social graph. Semantic Web (BERNERS-LEE *et al.* 2001) search engines such as Swoogle and Sindice search over structured data embedded in web documents, including data from Linked Open Data (LOD) and DbPedia (AUER *et al.* 2007).

Many of these search systems use keyword or natural language interfaces and analyze the query to infer latent entity structure. This includes identifying entities or their types, which is one step in what is broadly referred to as ‘query understanding’. Entity-based

query understanding has received significant recent attention. Aspects of research in this area include: automatically extracting structured attributes from queries (LI *et al.* 2009), extracting entity type classes (PASCA 2013), as well as mapping queries to entities in a knowledge base (MEIJ 2010). For example, the questions given to the IBM Watson DeepQA system, from “Jeopardy!” are automatically annotated with 2500 structured lexical answer types such as country, film, company, or author (FERRUCCI *et al.* 2010).

Recently, some of this annotation data has been made publicly available. The Google FACC1 data set (GABRILOVICH *et al.* 2013) is an open data set that includes queries (and web documents) annotated with entities from the Freebase knowledge base. Research on recognition and linking of entities in both documents and queries is an ongoing area of research in both the retrieval and natural language processing communities, with the Entity Recognition and Disambiguation (ERD) data challenge at SIGIR 2014¹.

Another trend driving increased use of entities in queries is interactive search interfaces that use auto-completion. As users type keywords, entity auto-suggestions appear and are added to the query, seamlessly subsuming keywords. For example, Facebook Graph Search² allows users to construct a structured query such as, [Photos of my family taken at national parks] that defines a type of return object (images), a social relationship (family member), and a place attribute (national park). The structured relationships between entities in databases support retrieval using inference and second order relationships impossible with text alone. For example the query [single malt scotch produced by distilleries founded before 1900] specifies a relationship between a distillery, a date, and a product.

¹<http://web-ngram.research.microsoft.com/erd2014/>

²<https://www.facebook.com/notes/facebook-engineering/under-the-hood-building-graph-search-beta/10151240856103920>

1.2 Beyond Entity Retrieval

Although search over only entities addresses many common needs, it does not include the broader context of these entities in text. One of the primary goals of this thesis is to effectively combine text and entity representations. A combined representation provides cross-document links from shared entity mentions, structured elements from mentions disambiguated to knowledge resources, as well as the scope and simplicity of text. We now highlight some of the limitations of focusing only on structured retrieval of entities.

The first limitation of entity retrieval is that entity databases are limited in size and scope. Entity databases vary widely in their freshness, accuracy of information, and completeness. The largest publicly available knowledge base is Freebase (BOLLACKER *et al.* 2008). It is a general purpose knowledge base containing 42 million entities and 2.3 billion facts³. Larger proprietary extensions of this exist with hundreds of millions of entities⁴. However, even for important relationships, a large fraction of important relationships are unknown: 68% of people in Freebase do not have a profession, 71% do not have a place of birth, and 91% do not have any education information (WEST *et al.* 2014). And although current general purpose knowledge bases have evolved significantly, the schemas are largely manually constructed. The result is that relevant entities or attributes may not exist (or be up to date) in the database.

Second, current entity databases contain primarily structured attributes. The knowledge bases focus primarily on encoding facts and relations to other entities. Most contain little or no textual representation. While facts are useful, text narratives provide important context for understanding relationships. The success of text knowledge resources, like Wikipedia, demonstrates the utility of rich text representations. The consequence of this is

³According to `freebase.com` as of January 24th, 2014

⁴<http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>

that information needs expressed as text may not match the entity, or the desired information may not exist in the knowledge base alone.

Third, entity databases often attempt to encode a single objective ‘reality’. Their goal is to represent a model of the world. However, for many topics the attributes or relationships may be controversial, subjective, or widely disputed. As a result, entities may have missing or incorrect information. In contrast, text collections contain subjective and opinionated documents with wide coverage of diverse opinions.

Lastly, and most importantly, many information needs cannot be easily be mapped to structured schemas in the knowledge base. Queries contain vague, abstract, or subjective language. For example, “What is the best place to retire?”. In particular, information needs often include a description of a process such as: “how can you...”, “factors that led to...”, “what role does...”, “how did user experience change over time as a result of...”, etc... The main source of information for these types of questions remains text.

As a result of this limitations, search over structured databases is not enough to satisfy many information needs completely. Text documents continue to be the primary search medium because they do not have these limitations. Text is the largest volume of information available and being created. Although the meaningful size of the web is unknown, it contains over 60 trillion pages ⁵ and there are over 500 million tweets per day ⁶. Many of these text documents contain rich explanatory narratives. They include descriptions of events and processes. Because there is no defined schema every aspect and topic is covered. And they are often subjective, describing the author’s opinions or perspective. As a result, this thesis focuses on enriching the representation of text with entity data.

⁵<https://www.google.com/insidesearch/howsearchworks/thestory/>

⁶<http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm>

1.3 Opportunity of Entity Representations

Entity-based representations address two fundamental issues with current word-based models: ambiguity and local uncertainty. Entity representations reduce ambiguity because disambiguated entities represent semantic units that exist in the world. The second issue is local uncertainty. Entities provide unambiguous cross-document links based on shared mentions of real-world concepts. This is important because not all information may be present in a single document. For an information need on [ferry sinkings with more than 100 people], one document may refer to a ferry sinking and another document may refer to its transport capacity, the joint mention of the same ferry entity allows cross-document inference.

The first issue that entity representations address is that words on their own are inherently ambiguous. A word may be polysemous and have many meanings. For example, a crane may refer to a bird, a type of construction equipment, a paper company, a welding company, or other meanings. Even for a relatively unambiguous name such as [Amherst], it may refer to a variety of entities and entity types including people [Jeffrey Amherst], locations ([Amherst, Massachusetts]), or organizations ([Amherst College]). Further, a word or name mentioned on its own may be missing important context, for example that a reference to the “Bounty” is a historical wooden ship. The result of this missing implicit information is the well-known problem of “vocabulary mismatch” (FURNAS *et al.* 1987).

Entity representations address this problem because they contain mentions of things, which are disambiguated to a knowledge base. This allows us to leverage text and structured attributes from the entity data. These features include both words and structured attributes. For example, the entity’s name (HMS Bounty), aliases (HM Armed Vessel Bounty), types (/boats/ship, /tallships/replica), and word distribution (from the knowledge base description or associated mentions), as well as other structured attributes relevant to the task. Some of the structured attributes might include: gender, nationality, profession, geographical

information (latitude, longitude), temporal attributes (such as birth and death), appearances in a movie, category classifications, etc.

The second problem addressed by entity representations is that local context may not provide sufficient information. In both information extraction and information retrieval, inference is typically performed one sentence or one document at a time. Local independence is a common assumption in most widely used information retrieval (ROBERTSON 1977) and information extraction algorithms. This view asserts that the observations in a document, D , are independent of those in other documents. However, this view is quite limiting. A single document or sentence in isolation may have sparse local information. Examining only local evidence leads to inconsistent labelling in information extraction and sub-optimal ranking in retrieval. Entities provide a mechanism to induce fine-grained dependencies based on shared mentions. Topically similar entity mentions provide links across sources. As we show, these long-distance edges are useful sources of features for local expansion and contextual weighting.

Entity-based enrichment has some of the following considerations:

1. For what task? (Document retrieval, question answering, in-document coreference, entity linking)
2. Entities from what source? (A knowledge base, mentions within the collection, mentions in external collections)
3. What types of features? (Words, aliases, fine-grained entity type, capitalization features, etc.)

We examine a variety of tasks, sources of entity information, and feature representations in this thesis.

1.4 Text to Entity Representations

In order to leverage entity information, it is necessary to identify and resolve entity mentions in text to a knowledge base. One means of adding entity data to text documents is automatic information extraction. The task of ‘entity linking’ to a knowledge base has recently received significant attention at the Text Analysis Conference (TAC) Knowledge Base Population (KBP) Entity Linking Task (JI *et al.* 2011). In this task mentions of traditional named entities (people, geo-political entities, and organizations) are linked to a knowledge base derived from Wikipedia infoboxes. There is also increasing interest in more general concept entities, with the task of ‘wikifying’ (RATINOV *et al.* 2011; HUANG *et al.* 2008; KAPTEIN *et al.* 2010) documents by linking them to Wikipedia.

Driven by increased search engine adoption, web content owners are increasingly embedding structured entity data into text documents in machine readable format. Two widely used schemes on the web are Facebook’s Open Graph Protocol⁷ and Schema.org⁸. A 2012 study showed that 30% of web documents contain some form of embedded structured data in RDFa or Microformats (MIKA and POTTER 2012). These structured attributes and relationships are already being used by commercial search engines to improve result presentation via rich snippets and support exploratory search through faceted browsing.

In 2013, Google released the publicly available FACC1 data set (GABRILOVICH *et al.* 2013) for the TREC ClueWeb09 and ClueWeb12 web collections. This data set contains automatically extracted entity mentions from web documents that are linked to the Freebase knowledge base (BOLLACKER *et al.* 2008). The FACC1 data set is the first openly available web-scale collection of entity linked documents.

In this work, we start from unstructured text documents and through information extraction build up increasingly sophisticated entity-based representations. At each step

⁷<http://ogp.me/>

⁸<http://schema.org/>

in the information extraction process we leverage our enrichment framework to leverage cross-document evidence from similar entities.

1.5 Entity-based Enrichment

The entity enrichment framework we describe in this dissertation is general enough to handle both diverse types of enrichment targets as well as different types of entities and their representation. Enrichment can be performed on words, entity mentions, documents, or even queries. Because of the rich structured attributes available in external knowledge sources, a significant focus of this thesis is on detecting, linking, and exploiting entity mentions that are linkable to these knowledge bases.

Our enrichment framework models cross-document dependencies in text documents based upon shared mentions of related entities. It builds a local context model and uses information retrieval as a mechanism to identify topically similar entity sources and their associated entity mentions. Entity enrichment differs from other cross-document entity models that treat all mentions of entities across documents equally, ignoring their similarity to another. This is important because even if two documents refer to the same real-world entity they may refer to different aspects of it. Lastly, retrieval is a highly scalable mechanism that allows us to identify similar entity mentions from unlabeled text documents in large external sources, such as the Web.

The enrichment process is related to other relevance feedback models, including relevance modeling (LAVRENKO and CROFT 2001) and latent concept expansion (METZLER and CROFT 2007; LANG *et al.* 2010). In these models the representation of the local query model is expanded using words and concepts from retrieved documents. In contrast to these approaches, our enrichment process is used to expand either document or query representations. It is focused on leveraging features from topically similar entity mentions. These features may include words, but are task-specific depending upon the document analysis

task. And because these entities may be disambiguated to a knowledge base we can leverage structured attributes from these mentions.

Entity Enrichment expands the feature representation of a local model with features from topically similar entity mentions in the document collection or external knowledge base. The enrichment process consists of several important steps which vary for each task and collection. The steps in entity-based enrichment are: 1) enrichment triggering, 2) target model generation, 3) mention retrieval, 4) mention feature extraction, and 5) feature aggregation. We now briefly describe this process for each of the tasks we study in this thesis.

1.6 Tasks

In this thesis we study applications of entity-based enrichment to three extraction and retrieval tasks: 1) Named Entity Recognition, 2) Entity Linking, and 3) Ad hoc Document Retrieval. These tasks build upon one another in levels of understanding documents through entities. The first task detects entity mentions, the second links entity mentions to external knowledge resources, and finally the third leverages the disambiguated mentions to improve retrieval effectiveness. We show how task-specific entity features are used for each of these tasks. We now provide an overview of enrichment in these tasks.

1.6.1 Named Entity Recognition

Named entity recognition is a pattern recognition task that assigns categorical entity labels (person, organization, location, miscellaneous) to a sequence of observed words. The goal is to infer a hidden label y from the observed token sequence x . The enrichment target in this task is the feature representation of each observed token, x_i . For mention retrieval, a query, Q_T is generated from the words in the sentence and sentence retrieval is performed on the text collection. Entity recognition features, such as adjacent words, part of speech tags, and Wikipedia gazetteer matches are extracted from string identical

observations in the retrieved sentences. The features are aggregated as we describe in more detail in Chapter 4. The result is both a local feature representation and one from similar mentions in other sources. We show that incorporating these non-local features results in entity recognition models that are more effective and more robust than local models. After entities are detected and classified, another further step is to disambiguate them.

1.6.2 Entity linking

Entity Linking is the task of disambiguating entity mentions, $m \in M_D$, in documents to entities, $e \in E$, that exist in a knowledge base. For example, the mention “Arran Distillery” should match its corresponding entity, [Arran_distillery] in Wikipedia. We model this as a retrieval task, where entities are ranked for each entity mention. A key factor in this process is identifying disambiguating context for the entity mention. For this task each mention m is a target for enrichment. Because the entire local document is the context, a fundamental problem here is defining the mention’s relationship to other entity mentions and words in the document. The main enrichment feature in linking is the strength of association, which we refer to as the salience, ρ , for words and entities in the contextual neighborhood of the target mention. This feature is used to generate expansion queries for retrieving entities in the knowledge base. Once entities have been detected and disambiguated, the next step is to leverage the disambiguated mentions to improve the effectiveness of other tasks.

1.6.3 Document Retrieval

The last task we explore in this thesis is entity-based enrichment models for ad hoc document retrieval. Our work addresses two fundamental research areas using entities for ad hoc retrieval. The first is the representation of both queries and documents with linked entities. What entity features, if any, improve retrieval effectiveness? The second is inferring latent entity-based query features for an information need. Linked entities provide a wealth of rich features that could be used for representation. These include both text as well as structured data. Some of the important attributes that we experiment with include:

fine-grained type information (athlete, museum, restaurant), category classifications, and associations to other entities. To model the context of linked entities we introduce novel query-specific entity context models extracted from snippets in the feedback documents surrounding the entity’s annotations.

1.7 Contributions

In this thesis we address the task of enriching the local representation using features from topically similar entity mentions across documents and in structured knowledge sources. Starting from text we build up increasingly sophisticated entity-based representations by automatically detecting and disambiguating entity mentions to external knowledge resources. We leverage the context and structured attributes from disambiguated entities to improve ad hoc document retrieval. We now detail the main contributions of this dissertation:

1. **We introduce a new expansion model, *entity enrichment* which performs feature expansion using topically similar entity mentions.** Unlike existing methods which perform expansion at the document level using words, our model expands the local feature representation using task-specific features from similar entity mentions across documents.
2. **We empirically demonstrate the effectiveness of entity-based enrichment for named entity recognition, named entity linking, and ad hoc document retrieval tasks.** We show that enrichment significantly improves the effectiveness over existing techniques for both information extraction and retrieval tasks. The results show a 6.8% error reduction on news wire and a 19.9% error reduction on out-of-domain book data for named entity recognition, up to a 16.4% improvement in mean reciprocal rank for entity linking, and gains up to 32.3% in mean average precision for ad hoc document retrieval.

3. **We define a new query-specific entity context model that models the feature context of an entity using retrieved documents.** Existing entity context models are built globally across a collection. Previous local models (XU and CROFT 1996) use word and phrase features from noun phrases that are not disambiguated for expansion. We introduce a new query-specific entity model that includes both words, phrases, and features from disambiguated entity mentions. For the task of retrieval, we show that these models provide an effective mechanism for identifying the relevance of entities and as a source of expansion features.
4. **We extend existing dependency models to include entity-based features that model dependencies between text and different types of entity features.** Existing query expansion techniques (METZLER and CROFT 2007) model dependencies derived from words, including phrase and proximity concepts. We propose a feature expansion model that models dependencies between text and structured entity features including: entities, fine-grained types, categories, and entity associations. For example a dependency between a type of entity and a word: (/boats/ship sinking) or (/government/politician scandal).
5. **We present the first known experimental results using entity linked documents and queries for ad hoc document retrieval.** We experiment using linked entities for newswire and web test collections. We use documents annotated with linked entities provided by the KB Bridge entity linking system and the openly available FACC1 entity annotations by Google for web data. For a subset of these collections we also experiment with entity linked queries. We compare models incorporating entity features with state-of-the-art word-based models. Compared with competitive query expansion baselines, the sequential dependence model with relevance modeling expansion on Wikipedia and the collection, there is an improvement of 16.4% and 11.5% in MAP on Robust04 and a 14.1% and 32.8% improvement in NDCG@20 for

ClueWeb12. For ClueWeb09, where results do not significantly improve, we perform an error analysis and identify several important underlying causes for this behavior.

1.7.1 Outline

- Chapter 2 surveys related work, and provides background used in the entity-enrichment model.
- Chapter 3 describes the data collections, evaluation methods, and metrics used in this dissertation.
- Chapter 4 introduces the entity enrichment model (Contribution 1) used for each of the tasks described in this thesis.
- Chapter 5 presents enrichment applied to the task of named entity recognition. It presents experiments evaluating the utility of leveraging cross-document features from similar entity mentions for newswire and book collections (Contribution 2).
- Chapter 6 presents enrichment applied to the task of entity linking. It details an investigation of a retrieval-based approach for linking mentions to an external knowledge base. This chapter is the second application of the enrichment model for identifying disambiguating context for entity linking. It presents empirical results investigating the effectiveness of various contextual components (Contribution 2).
- Chapter 7 investigates entity-based feature expansion for ad hoc document. It explores structured expansion on queries incorporating feature dependencies (Contribution 4). One source of expansion is query-specific entity context models (Contribution 3). It presents the first results evaluating the effectiveness of features from linked entities for document retrieval (Contribution 2 and Contribution 5).
- Chapter 8 summarizes the contributions made in this thesis and discusses possible future research directions in the area.

CHAPTER 2

BACKGROUND

In this chapter, we describe the background and related work. Entity-based enrichment incorporates several different threads of research in information retrieval. Entity enrichment is a type of structured expansion of the local representation using entities. It is related to previous work on query and document expansion that expand the local representation by adding new words and re-weighting existing ones. The enrichment model in this thesis focuses on entities, particularly those disambiguated to an external knowledge base. This is related to previous work using structured concept vocabularies in retrieval. It is also related to previous work on disambiguating word senses for information retrieval.

We conclude this chapter with a description of the retrieval models we build upon throughout this thesis. Graphical Models (KOLLER and FRIEDMAN 2009) are used in both information extraction and information retrieval tasks. As a result, we use them as a common representation for the models described in this work. For retrieval, one of the main models we use is the Markov Random Field (MRF) retrieval framework both for modeling dependencies in both the original and expansion queries.

The related work presented here is relevant to the enrichment framework overall. We also present other related work in the corresponding chapters as needed.

2.1 Query Expansion

Query expansion is a type of query transformation where the original query is augmented by adding terms and possibly reweighting terms. Expansion is used to address the problem of vocabulary mismatch (FURNAS *et al.* 1987) between the query and documents. Query

expansion has a long history in information retrieval (ROCCHIO 1971; CROFT and HARPER 1979; SALTON and BUCKLEY 1990). These include global approaches, which leverage collection-wide term clusters based on word co-occurrences (JONES and BARBER 1971). An alternative approach is a query-specific approach that uses top-ranked retrieved documents for local feedback (ATTAR and FRAENKEL 1977; CROFT and HARPER 1979; BUCKLEY). These techniques are referred to as relevance feedback and pseudo relevance feedback (PRF). Pseudo relevance feedback is an unsupervised automatic expansion technique that leverages evidence in the top retrieved documents. Harper and Croft (1979) use pseudo relevance feedback to reweight the original query terms. We use a similar approach in this thesis to reweight relationships between entities for entity linking in Chapter 6.

Pseudo relevance feedback is an area of research that has received significant attention (LAVRENKO and CROFT 2001; DIAZ and METZLER 2006; LV and ZHAI 2010; ZHAI and LAFFERTY 2001; BENDERSKY *et al.* 2012). One widely used expansion model is the relevance model (LAVRENKO and CROFT 2001). It is a pseudo-relevance feedback approach that uses retrieved documents to estimate the query topic. Relevant words are extracted and used in combination with the original query (RM3). Throughout this thesis we use this as our baseline text-based expansion model.

Another feedback model that incorporates term dependencies in addition to words is the latent concept expansion (LCE) model proposed by Metzler and Croft (2007). It builds upon the Markov Random Field retrieval framework (METZLER and CROFT 2005) and introduces the idea of using arbitrary features for expansion. However, in their experiments they use only unigram text features because they find others do not significantly improve retrieval effectiveness. The use of named entities as expansion concepts was examined by Abdul-Jaleel *et al.* (JALEEL *et al.* 2005), who experimented with named entities as expansion terms for the TREC HARD task, but the results were inconclusive. They used uniform weighting of entities, and highlight the potential need for more effective weighting methods. In this thesis, we use entities and features derived from entities as conceptual units,

which goes beyond the previously used word, phrase, and proximity concepts previously proposed. In Chapter 7 we also extend the work to incorporate dependencies between words, mentions of entities, and features of linked entities. Unlike their work, the evaluation we perform shows significant improvements in retrieval effectiveness from these concepts.

The Phrasefinder (JING and CROFT 1994) approach performs query expansion using words or phrases by building a global context model from the source collection. The context of a word is an aggregation of word co-occurrence counts within a paragraph. This constructs a pseudo-document for a phrase that can be indexed. The user query is then issued against this index to find related phrases. They find that noun phrases were the most effective for expansion. It is outperformed by methods (XU and CROFT 1996) that build contextual models of noun phrases specific to the query from retrieved documents. As a result, in this thesis we focus on local models derived from similar entity mentions.

Another related pseudo relevance feedback model using expansion ‘concepts’ is Local Concept Analysis from Xu and Croft (1996, 2000). Local context analysis identifies expansion ‘concepts’, nouns and noun phrases, from top retrieved documents. It uses unigrams and phrase word features based on co-occurrences near query terms in top ranked documents. The contribution of words versus phrase concepts was not evaluated. The examples show that many of the expansion terms appear to be single words. In contrast, this work focuses on entities, predominantly named entities. One key difference in this work is that we leverage disambiguated entity mentions linked to a knowledge base. Instead of the terms themselves, we extract features from the mentions.

Wikipedia as a source of world knowledge has been demonstrated to improve a variety of tasks, including retrieval. It has been demonstrated to improve retrieval effectiveness when used as an external source of terms for query expansion (DIAZ and METZLER 2006; BENDERSKY *et al.* 2012; XU *et al.* 2009). All of these previous methods treat Wikipedia as a text corpus. Some leverage superficial metadata, such as redirects. In contrast, this work uses Wikipedia as a knowledge base of entities. We link text documents to Wikipedia and use

entity features from it for feature expansion. In related work, Meij et al. (2009, 2010) map search engine queries to DbPedia using supervised machine learning and use the concepts for query expansion, showing small improvements in effectiveness and improved result diversity. Balog et al. (2011) integrate term-based and category-based query representations for entity retrieval and find that category-based feedback is more effective than term feedback with both pseudo-relevance feedback and relevance feedback. All of these representations focus on models of the query rather than the document representation.

2.2 Document Expansion

In this thesis we enrich document feature representations as well queries. For information extraction tasks a query is not generally available. Enrichment on documents is closely related to previous work on document expansion. Similar to query expansion, document expansion adds terms or term weights to the representation of a document.

Document expansion models attempt to solve sparsity and insufficient sampling problems, particularly for short documents. Document expansion identifies similar documents based on similar word usage (BLAIR 1979) as well as similar structural citation patterns (CROFT *et al.* 1989). Liu and Croft (2004) propose cluster based language models to smooth a document model. Tao et al. (2006b) define a document neighborhood using cosine similarity, and use this to construct an *enlarged* document mode. Mei et al. (2008) propose an optimization framework for smoothing language models on graph structures, with the goals of providing *fidelity* and *smoothness*. Diaz (2008) proposes a query-specific model that regularizes retrieval scores based upon document similarity.

For tweets, Efron et al. (2012) perform document expansion using the relevance modeling framework, treating the tweet as a query and interpolating the document language model with retrieved documents. One key finding of their work for microblog retrieval is that expanding the query performs poorly, but expanding the documents results in consistent effectiveness improvements.

Singhal and Pereira (1999) perform document expansion for speech retrieval over audio transcribed from automatic speech recognition (ASR). They find that document expansion is very effective in improving poor transcriptions. They perform expansion using Rocchio's method (ROCCHIO 1971) on the 10 closest documents. They issue the entire ASR document as a query against a collection of newswire documents to find the nearest neighbors. They find that there are significant gains from reweighting the terms (17%) in the document and from adding new terms (6%) for short queries.

Metzler et al. (2009) and Yi and Allan (2010) propose methods for overcoming anchor text sparsity in web search by enriching the document representation with text that is aggregated across the hyperlink graph based on shared links and content similarity.

All of these methods perform document expansion using words. In contrast, in our work we enrich document representations with task-specific features including words, but more importantly with other semantic features. Instead of clustering documents, we focus on identifying similar entity mentions. Expansion is performed from these mentions rather than the document overall. This is a more fine-grained expansion approach than previous models.

2.3 Word Sense Disambiguation

Because words are ambiguous there have been efforts to represent documents using other types of representations. The first area of research is Word Sense Disambiguation (WSD) which indexes disambiguated sense of individual words.

Work on indexing word senses in information retrieval received significant attention in the late 1980s and 1990s. Krovetz and Croft (1989, 1992) studied word sense disambiguation in the context of retrieval. They find that lexical ambiguity is not a significant problem in documents when there is a match to multiple query words. They highlight that a more important factor may be particular relationships between words. Voorhees studied this problem where each noun was disambiguated to Wordnet (VOORHEES 1999). On their own, they found that using senses degraded retrieval effectiveness because of an increased number

of mismatches between query terms and documents. Sanderson found that disambiguation accuracy of at least 90% was needed to avoid hurting effectiveness (SANDERSON 1994). Mihalcea and Moldovan (2000) find modest improvements using disambiguated word senses and find that only 55% of words can be disambiguated. One well-documented issue with WordNet is its coverage, particularly for proper names (SCHÜTZE and PEDERSEN 1995).

In contrast to these approaches which focus on disambiguating words, we focus on entities. These entities may be disambiguated to a knowledge base. Using entities instead of words is important because they form units of meaning used across documents. They have attributes and structured relationships to other entities in the knowledge base and other documents. Beyond decreased ambiguity, entities provide richer representations than simply words. For example, they may be linked to specific geographic and temporal scope. Similar to previous work on WSD, the ability to detect and disambiguate entities correctly is one factor limiting their utility for retrieval applications. As we show in Chapter 7 there remain significant mismatch and detection gaps in current entity detection and linking systems. For this reason, this thesis combines both text and entity-based representations.

2.4 Controlled Vocabulary Representations

The use of structured knowledge resources has a long history in the field of information retrieval and before that in library science. Early work in digital systems is an extension of library science where works are classified using index terms from controlled vocabularies, like the Universal Decimal Classification, by professional librarians. For example, documents were indexed using physical punch cards or metal plates and retrieved using rods which found cards with holes in the correct locations (JOYCE and NEEDHAM 1958).

2.4.1 Structured Classification

In the era of digital systems, the use of manually assigned index terms as a means of representing documents continued. Early work using these in digital domains in the

1950s include Uniterms, Zatacoding, and thesaurii (LUHN 1957; JOYCE and NEEDHAM 1958; SALTON 1968) to assign index terms to documents. This continued in specialized domains. In the 1960s the first computerized medical library systems adopted these methods. The Medical Literature Analysis and Retrieval System (MEDLARS) from the National Library of Medicine created the Medical Subject Heading (MESH) controlled vocabulary (LIPSCOMB 2000). These systems continue to be used today for search today, including in the MEDLINE and PubMed collections and have evolved to include other vocabularies, such as the Unified Medical Language System (UMLS) metathesaurus. Lin and Demner-Fushman (2006) demonstrate that combining these structured concept vocabularies with text for search in the domain of clinical medicine results in significant retrieval effectiveness gains.

The use of manual topical classification of content continued with the advent of the Internet. Manually curated web directories became popular in the 1990s and included the Yahoo! directory and the Open Directory Project (ODP). Gauch et al. (2003) map documents to topic models derived from the Open Directory Project (ODP), where documents are categorized via text classification. Wei and Croft (2007) evaluate ODP category models for retrieval and find that they improve queries that lack a clear topic, but are outperformed by relevance models when the topic is specific and clear. They hypothesize that the categories may be too broad for some information needs.

However, using manually assigned index terms as the primary means of document representation has been repeatedly demonstrated to be inferior to full-text retrieval. The use of full-text versus index terms as a means of document representation was the focus of the early Cranfield experiments (CLEVERDON 1991). It was also studied in medicine using the SMART system comparing it to MEDLARS (SALTON 1972). On the web, the rise of full-text search engines including Google and Bing have replaced directories.

Two other controlled vocabularies that have been widely used are Cyc and Wordnet. Cyc¹ is a machine readable ontology of commonsense knowledge (LENAT 1995). It has been used in question answering, but had minimal impact (CHU-CARROLL *et al.* 2003) because of lack of scope. Wordnet (MILLER 1995) is a lexical database of English which arranges words into equivalence classes, called synsets which represents the sense of a word. It also models relationships between synsets. Wordnet has been used by a variety of researchers to perform ‘semantic indexing’, based on word senses. It suffers from similar gaps in coverage. Wordnet focuses on words, but does not include named entities.

The rise of user-generated content in the early 2000s, including Flickr and Delicious, gave rise to informal user generated ‘tags’ as a form of ‘folksonomy’ (GOLDER and HUBERMAN 2006), an informal ontology. These methods and earlier ones assign document-level metadata ‘tags’ consisting of index terms from structured vocabularies. In contrast, we focus on annotations at the level of individual mentions that occur within documents.

2.4.2 Wikipedia

Wikipedia has been used as means of representing documents for a variety of NLP tasks, which is sometimes referred to as Explicit Semantic Analysis (ESA). Explicit Semantic Analysis represents each word as a vector of the most relevant Wikipedia articles. It has been shown to improve text categorization (GABRILOVICH and MARKOVITCH 2006), semantic relatedness (GABRILOVICH and MARKOVITCH 2007), and document clustering (GABRILOVICH and MARKOVITCH 2007). For retrieval, Egozi *et al.* (2008) use ESA concepts to augment the traditional word based document representations. They evaluate on one small collection, TREC-8, and show some improvements (EGOZI *et al.* 2008), between 4-15%. They use pseudo-relevance feedback from ESA-annotated text documents to identify concepts and also experiment with fusing text and concept-based scores. One issue with ESA representations is because mapping is done at the word level, the semantics of phrases and

¹<http://www.opencyc.org>

larger entities may be lost. In contrast, in this work we focus not on words, but on identified entities as units of representation which we detect and optionally link to a knowledge base. In contrast to representing a document as a vector of Wikipedia concepts, we leverage the associated entity representation of entities from the knowledge in the form of their text and structured metadata to enrich the original document representation.

2.4.3 Implicit Topics

Another mechanism for addressing the underlying issues of sparsity and ambiguity in local representations is a different representation of documents. Implicit concepts are typically lower dimensional representations of observed terms. Previous work using them includes latent semantic indexing (DUMAIS 1995), probabilistic latent semantic analysis (HOFMANN 2001), and latent dirichlet allocation (BLEI *et al.* 2003; WEI and CROFT 2006), and restricted boltzmann machines (WELLING *et al.* 2004). For retrieval, the use of topic modeling was examined recently by Yi and Allan (2009). They found that topic models did not perform well and was outperformed by relevance modeling (LAVRENKO and CROFT 2001). They found only small improvements when topic models were combined with relevance modeling. The reason for this is that the lower dimensional representations rarely match the granularity of users' information needs. In contrast, we focus on explicit entity concepts because these are finer-grained units of representation that people have deemed noteworthy by creating a knowledge base entry with facts and relationships.

2.5 Entities in Retrieval

The research area of retrieving entities as well as using named entities has received significant attention. It has been studied at a variety of venues. The TREC entity retrieval track (BALOG *et al.* 2011) focused on entity-oriented search tasks and ran from 2009 through 2011. The tasks include related entity finding, e.g. airlines that use the Boeing 747 airplane, from both webpages as well as Linked Open Data (LOD) collections. Similarly, the INEX

entity ranking (DEMARTINI *et al.* 2010) track focused on retrieving entities from Wikipedia, e.g. art museums in the Netherlands. The INEX Linked Data track (WANG *et al.* 2011) also focused on Wikipedia, but also explored retrieval and ranking over additional structured data in RDF format. Retrieving entities has also been the focus of several workshops, including the Workshop on Entity Oriented and Semantic Search (BLANCO *et al.* 2011; BALOG *et al.* 2012) and the Semantic Search Workshop (TRAN *et al.* 2010; TRAN *et al.* 2011). In SemSearch 2010 we examined the retrieval models that we use in this work and found that they were an effective technique on structured RDF entity data as well (DALTON and HUSTON 2010).

The TREC Enterprise Track (CRASWELL *et al.* 2005; BALOG *et al.* 2008) ran from 2005 through 2008 and includes the expert search task, with the goal of ranking people within an organization who are subject experts in a particular field. One widely used approach for modeling people is building a profile of the entity from the text (PETKOVA and CROFT 2007) and by exploiting relationships between the relevance of documents and the people mentioned within them (BALOG and DE RIJKE 2008).

Topic Detection and Tracking (TDT) (CONNELL *et al.* 2004) is a research program that ran for seven years from 1998 through 2004. The program investigated methods for organizing news articles as they arrive in a stream. The tasks include identifying the appearance of an event and tracking their evolution over time. There were several approaches leveraging entities to model events. Kumaran and Allan (2004) study New Event Detection (NED) and utilize detected named entity strings as a representation of a document. They highlight the importance of entity representations, but focus only on the word representation of entities. Similarly, entities were also shown to be effective document representation for Story Link Detection (SLD) (SHAH *et al.* 2006). We similarly find entities to be an effective feature in representing local context. Beyond the names of the entities themselves, we focus on detecting and leveraging entities that exist in structured and semi-structured knowledge bases.

The TREC Knowledge Base Acceleration (KBA) track (FRANK *et al.* 2012) is another news stream filtering track, focused on entities. In this task, entities are the primary unit of interest and the tasks for the track involve identifying “vital” documents with timely and new information to update a knowledge base entry. A second task focuses on streaming slot filling, where the goal is to fill in attributes of an entity, such as aliases or a birth date. Although it is not the main focus of this work, we have applied the entity context modeling and entity linking approaches used in this thesis to this task (DALTON and DIETZ 2012).

In contrast, we focus on document retrieval leveraging automatic entity annotations. Exploiting entity links and other types of semantic annotations is an open area of research. The workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR) (BENNETT *et al.* 2013; KAMPS *et al.* 2012; KAMPS *et al.* 2011) has run over the last five years, and highlights the need for continued research in this area.

As part of the TIPSTER project and TREC-2 conference, the INQUERY retrieval system incorporated basic entity information (CALLAN *et al.* 1994). Entity mentions are recognized using sequences of capitalized words. Then gazetteers are used to classify entities as company names, U.S. cities, and countries. of companies and foreign countries were identified using capitalized sequences of words classified using simple gazetteers. They demonstrate small improvements in precision by careful use of these entity concepts.

2.6 Graphical models for IR

Graphical models (KOLLER and FRIEDMAN 2009), such as Markov Random Fields (MRF) (METZLER and CROFT 2005), are a popular tool in both information extraction and information retrieval. After casting data and quantities of interest as random variables, dependencies between two (or more) variables are encoded by factor functions ϕ that assign a non-negative score to each combination of variable settings. Factors ϕ are often expressed by a log-linear function of a feature vector. The joint configuration of all variables is scored by the likelihood function \mathcal{L} , which is represented by the normalized product over scores

of all factor functions for the given variable settings. We distinguish notationally between random variables in upper case (e.g. E) and possible assignments (e) in lower case. We denote count statistics of a configuration e in a sequence as e_i .

2.6.1 Log-linear Models

Dependencies between two (or more) variables (e.g. queries and documents) are encoded by factor functions that assign a non-negative score to each combination of variable settings. Factor functions (or similarity function) between two variables are indicated by ϕ (e.g. $\phi(Q, W)$) which is assumed to be of log-linear form. This means that ϕ is determined by an inner product of weight vector θ and feature vector f in log-space.

2.6.2 Retrieval Models

The query likelihood (QL) retrieval model can be represented as a factor between a query consisting of multiple query words, and a document represented as a bag of words as

$$\phi(Q, D) = \prod_{w_i \in Q} \phi(w_i, D)$$

The feature function used to match words, W , to a document is a Dirichlet smoothed probability:

$$\phi(W, D) = \log \frac{\#(W, D) + \mu \frac{\#(W, C)}{|C|}}{|D| + \mu} \quad (2.1)$$

Within this framework, we use the bag-of-words query likelihood model as well as more recent models that capture term dependence.

One of the most widely used models that captures dependence relationships is the sequential dependence model (SDM) (METZLER and CROFT 2005), which incorporates word unigrams, adjacent word bigrams, and adjacent word proximity. The key dependencies

it captures are adjacent terms. The score for these ordered bigrams (represented as the function #1 from the Galago query language) is:

$$\log \phi^{\text{O}}(q_i, q_{i+1}, d) = \theta^{\text{O}} \cdot \#1(q_i, q_{i+1})$$

and similarly for unordered bigrams within a window of eight terms (represented as the function #uw8 in the Galago syntax) as:

$$\log \phi^{\text{U}}(q_i, q_{i+1}, d) = \theta^{\text{U}} \cdot \#uw8(q_i, q_{i+1})$$

We use the sequential dependence model for representing dependencies from different kinds of feature vocabularies, such as entity identifiers or entity categories with appropriate redefinition of the document length $|D|$ and collection statistics.

2.6.3 Relevance Feedback

In both relevance modeling (LAVRENKO and CROFT 2001) and latent concept expansion (METZLER and CROFT 2007) the query expansion formulation from top retrieved documents is similar. Assuming that the retrieval score represents the probability of the document under the query, e.g. $p(D|Q)$, document-wise multinomial distributions over a vocabulary $p(V|D)$ are combined via a mixture model.

$$p(V|Q) = \sum_{D \in C} p(V|D)p(D|Q) \tag{2.2}$$

Hyperparameters of this approach are the number of expansion documents, number of expansion features, and a balance parameter for weighting the original query against the expanded query, which are further weighted according to $P(V|Q)$.

The document probability, $p(D|Q)$ is typically derived from the retrieval score $s(D)$ by exponentiation and re-normalization over the domain of expansion documents, R . The document specific distribution of features is derived under the multinomial assumption by

$$p(V|D) = \frac{\#(V \in D)}{\sum_{V'} \#(V' \in D)}$$

The result distribution over the features, V , from the collection. It is focused on documents which are topically related to the query. This is the framework that we use to generate focused models for entity-enrichment. We describe this further in Chapter 4.

CHAPTER 3

DATA

In the following chapters we will examine the task of entity-enrichment across three different tasks: named entity recognition, entity linking to a knowledge base, and ad hoc document retrieval. In this chapter we detail the experimental data used for each evaluation.

3.1 Named Entity Recognition Corpora

We use two test collections for evaluating our named entity recognition effectiveness. Our primary data set is a standard corpora from the Conference on Natural Language Learning (CoNLL) 2003 shared task. The second is the Deerfield Collection, a collection we construct from publicly available scanned and OCRed books on topic of the history of Deerfield, Massachusetts. Statistics for both test collections are shown in Table 3.1.

3.1.1 CoNLL 2003

The CoNLL 2003 English data set is a widely used collection created for the shared task of the Seventh Conference on CoNLL, which focused on entity recognition. It consists of news wire documents from the Reuters RCV1 TREC corpus from the year 1996. It is

	Deerfield	CoNLL Test
Tokens	10,050	46,435
Person	273	1617
Miscellaneous	98	702
Location	241	1668
Organization	49	1661

Table 3.1: NER collection statistics

annotated with four types of named entities: persons (PER), locations (LOC), organizations (ORG), and miscellaneous (MISC) entities. The data consists of three files: training, testa, and testb. As is commonly done, we train on a combination of training and testa, and report evaluation on the testb set. The combined training set (training and testa) contains 945 documents from August 1996 with 14,987 sentences and approximately 200,000 tokens. The evaluation testb set contains 231 documents from December 1996 with 3,584 sentences and approximately 46,000 tokens. We also note that documents in the collection are ordered by their appearance in the news stream, reflecting a temporal ordering.

3.1.2 Deerfield Book Collection

We created a named entity test collection using public domain books relevant to the history of the town of Deerfield, Massachusetts. The books are scanned by the Internet Archive ¹ and processed with OCR software to produce text. The Historic Deerfield collection contains ten books with 3,311 pages, 98,444 sentences, 2.1 million tokens, and over 60 thousand distinct words. It has diverse historical genres: biographies, encyclopedias, and historical catalogues of artifacts. To create a evaluation set for NER, we randomly sampled two pages from each book in the collection. The resulting test set contains 20 pages with 481 sentences and approximately 10 thousand tokens. The pages were manually annotated with entities consistent with the CoNLL task. ² The dataset contains 661 entity mentions.

3.2 Entity Linking

We base our experimental evaluation on data from the TAC KBP English entity linking competition from 2009 to 2012. The TAC data contains three elements: source documents,

¹<http://www.archive.org/details/texts>

²The collection and judgments are publicly available at <http://ciir.cs.umass.edu/~jddalton/deerfield>

Type	Documents
Broadcast Conversation	17
Broadcast news	665
Conversation Telephone	1
Newswire	2,286,866
Web	1,490,595

Table 3.2: TAC Source Corpus

Entity Type	Frequency
Person	114,523
Geopolitical	116,498
Organization	55,813
Unknown	531,907

Table 3.3: TAC Knowledge Base Types

Type	2009	2010 web	2010 eval	2011	2012
Person	627	500	741	750	919
Geopolitical	567	500	749	750	604
Organization	2710	500	750	750	706

Table 3.4: TAC Mention Types by Year

entity mentions to link, and a target knowledge base. These resources were developed by the Linguistic Data Consortium (LDC) for the evaluation (ELLIS *et al.* 2011).

The TAC source collection has evolved over time, adding new documents and mentions each year. The documents are drawn from a variety of other existing test collections including: ACE08, Gigaword, and GALE. For this work we use the TAC KBP 2012 English corpus, which includes the documents and mentions from 2009 through 2012. The statistics for the corpus are given in Table 3.2. It contains a heterogeneous mixture of content types including newswire, web documents, and a small quantity of transcribed speech.

The TAC reference knowledge base contains 818,741 entries from an October 2008 dump of English Wikipedia. A breakdown of the knowledge base by type of entity is given in Table 3.3. The knowledge base itself contains semistructured data in XML which was extracted from the infoboxes.

The entity linking task requires systems to link named entity mentions of persons (PER), organizations (ORG), and geopolitical entities (GPE) to the single coreferent entry in the knowledge base. This implies that there is exactly one relevant entity from the knowledge base for each target mention. If the mention does not have an entity in the knowledge base the system should detect this, label the mention as “NIL”, and perform cross-document coreference resolution to cluster the mention with other query mentions in the evaluation set. For the years 2009-2011 the query mentions were detected using an English named entity tagger. In 2012, annotators used a new annotation tool and were able to select arbitrary text extents. A sample of the mentions in the corpus are manually selected by LDC annotators as queries, with a bias towards highly confusable entity mentions including ambiguous names, misspellings, nicknames, and common names. For 2010 through 2012 the goal was to provide roughly an even distribution of the three entity types. The distribution of these types across years is given in Table 3.4. One interesting note is that the 2009 data contains a significantly larger number of mentions, with a significantly different proportion of organizations. For the years 2010-2012, approximately 2/3 of the mentions are from

newswire documents and 1/3 are from the web or informal documents. There is also roughly a balanced proportion of “NIL” and “In-KB” mentions.

3.3 Ad Hoc Retrieval

The study of entity enrichment for ad hoc retrieval uses data from TREC³ test collections for the experiments. We use the Robust04, ClueWeb09, and ClueWeb12 collections. These collections are from the 2004 Robust track and the Web track (from 2009-2013). Each collection consists of a text documents, a set of topics, and relevance assessments for the topics. A summary of the collections and topics used in this dissertation is shown in Table 3.5.

3.3.1 Documents

The definition of a *document* in TREC corpora varies widely. A document could be a news article, a book, a web page, an academic publication, emails, and even tweets in the microblog track. For the collections used in this dissertation, the documents are newswire articles and web pages. Both types of documents are text with some semistructured elements (such as a headline or title). The ClueWeb collections both contain two subcollections, Category-A and Category-B. Category-A is the complete collection. Category-B is a subset containing roughly 50 million documents, which is approximately 5-10% of the overall collection. For ClueWeb09, the Category-B subcollection consists of the web pages with the highest crawl priority (PageRank) pages as well as a snapshot of the English Wikipedia. For the ClueWeb09 collection we employ the Waterloo spam classifications (CORMACK *et al.* 2011) and filter the collection to documents that are in the 60th score percentile. For ClueWeb12-B, the documents are a uniform 7% sample of the 733 million documents created by taking every 14th document from the full Category-A set.

³<http://trec.nist.gov/>

3.3.2 Topics

Each TREC collection comes with a corresponding set of *topics* which represent information needs. An example of a topic is shown in Figure 3.1. Each topic has a short keyword title and a longer description. Only the titles were used in the original TREC track and we follow that convention. We use all 250 of the Robust04 topics as queries. The topics for ClueWeb09 are from the 2009-2012 TREC web tracks. The topics for ClueWeb12 are from the 2013 web track.

3.3.3 Relevance Assessments

For each topic the collection contains relevance assessments for a set of documents. The relevance judgments are performed manually by TREC assessors. For the Robust04 collection, binary relevance judgments are provided (relevant or non-relevant). For the web collection, the documents are rated on a graded relevance scale. The non-relevant grades are: (-2, junk), (0, non-relevant), (1, relevant), (2, highly relevant), and (3, authoritative). For binary relevance evaluation measures, we consider the first two (-2,0) to be non-relevant and the others relevant.

3.3.4 Entity Annotations

We experiment using linked entities in these documents for retrieval. For this data, we use two types of entity annotations. For the newswire collections, we create our own entity annotations. For both ClueWeb collections we use the Google FACC1 entity annotations (GABRILOVICH *et al.* 2013). The FACC1 dataset is the first publicly available web-scale collection of entity linked documents.

For the Robust newswire collection, we use our own entity annotations. For analyzing the documents, we use the NLP tools in the *factorie* (MCCALLUM *et al.* 2009) toolkit. We use *factorie* to process the documents and perform tokenization, sentence segmentation, named entity recognition, part-of-speech tagging, dependency parsing, and entity mention finding. The mentions detected by *Factorie* are linked to Wikipedia using the entity linking

Name	Documents	Topic Numbers
Robust04	528,155	301-450, 601-700
ClueWeb09-B	50,220,423	1-200
ClueWeb12-B	52,343,021	1-50
ClueWeb12-A	733,019,372	1-50

Table 3.5: Test Collections Statistics

described in Chapter 6. We note that these entity links are traditional named entities (people, organization, and geo-political entities) used in the TAC KBP evaluation.

For the ClueWeb collections, we use the publicly available entity annotations linking entities to the Freebase knowledge base provided by Google in the FACC1 dataset (GABRILOVICH *et al.* 2013). The dataset contains automatically extracted entity mentions that are linkable to the Freebase knowledge base (BOLLACKER *et al.* 2008). Freebase is a publicly available general purpose knowledge base with over 42 million entities and over 2.3 billion facts.⁴ Summary statistics for the annotations of both ClueWeb collections are shown in Table 3.6. An example annotation is shown in Figure 3.2. Google does not provide details on how the mentions are detected or linked. Only entity mentions that are linkable to the Freebase knowledge base with high precision are provided. The authors state that the precision is believed to be 80-85% and recall is estimated to be 70-85%.

In addition to the document annotations, the Google FACC1 dataset provides explicit entity annotations for the web track queries (2009-2012) for ClueWeb09. These are created by entity linking text in description field. We also experiment with a revised version of these annotations which improves recall and fixes several annotation errors. We discuss this further in Chapter 7.

⁴As of January 27, 2014 according to Freebase.com

```

<Id> 643
<Title> salmon dams pacific northwest
<Description> What harm have power dams in the pacific northwest
                caused to salmon fisheries?

```

Figure 3.1: Example Robust Query

Collection	Docs with Ann.	#mentions	avg / doc
ClueWeb09	340,451,982	5,107,067,522	15
ClueWeb12	456,498,584	6,133,750,307	13

Table 3.6: FACC1 ClueWeb Annotation Statistics

3.4 Evaluation

Each of the three tasks we study in this thesis uses its own evaluation measures. Common across all of these are precision, recall, and accuracy. Precision is the fraction of relevant (true positives) responses of all the responses returned:

$$P = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Recall is the fraction of relevant (true positive) responses of the total true results found by the system:

$$R = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Accuracy is the proportion of correct results:

$$A = \frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{false positives} + \text{true negatives} + \text{false negatives}}$$

Document	Mention	Start byte	End byte	Posterior	Freebase ID
clueweb09-en0000-00-04720	FDA	21303	21306	0.9998	/m/032mx
clueweb09-en0000-00-00005	G e	9188	9196	1.0000	/m/03bnb

Figure 3.2: Example FACC1 Entity Annotations

We now examine the evaluation measures for each task in more detail.

3.4.1 Named Entity Recognition

Named entity recognition is a sequence labeling task. In this thesis we follow the CoNLL convention (KIM *et al.* 2003), where each entity is correct only if the identified entity exactly matches the entity in the manual judgments.

We evaluate named entity recognition with the widely used F1 measure, the harmonic mean of precision and recall. It is a widely used variant of the F-Measure (BLAIR 1979) with $\beta = 1$. These measures are defined as:

$$F_{\beta} = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R}$$

$$F_1 = \frac{2 * P * R}{P + R}$$

For significance testing we use a random permutation test. To measure the difference between two output sequences, we run a Monte-Carlo permutation test using 1000 samples. If $\alpha < 0.05$ we conclude there is a significant difference between the systems.

3.4.2 Entity Linking

For entity linking, we evaluate both overall effectiveness as well as “In-KB”. For “In-KB”, there is only one correct entity and the relevance is binary (0 or 1). We evaluate this case as a ranked retrieval where there is one relevant result. For this, we can evaluate precision at rank 1 (P@1) as well as mean reciprocal rank (MRR). The P@1 is the average of the precision at rank 1 over a set of queries. The *reciprocal rank* measure is defined as the reciprocal of the rank of the first relevant retrieved document. The *mean reciprocal rank* is the average of the reciprocal ranks for a set of queries.

The primary evaluation measures for entity linking are based on the B-Cubed cluster evaluation measures (BAGGA and BALDWIN 1998). B-Cubed is a coreference evaluation

measure for scoring coreference clusters. It defines cluster precision and recall for each entity, i , as:

$$P_i = \frac{\text{number of correct elements in the output chain containing entity}_i}{\text{number of elements in the output chain containing entity}_i}$$

$$R_i = \frac{\text{number of correct elements in the output chain containing entity}_i}{\text{number of elements in the truth chain containing entity}_i}$$

As formulated the chains must only be consistent within the collection. It does not handle the case where the mentions have an identical grouping, but refer to different entities in an external knowledge base. For entity linking in TAC, these measures were modified to fix this issue. TAC refers to these measures plus variations, such as B-Cubed+. B-Cubed+ is the same as B-Cubed, but adds the constraint that the clusters must also be linked to the same entry in the external knowledge base. B-Cubed also defines two options for weighting elements in the evaluation: 1) cluster precision where each cluster is weighted equally and within the cluster each query is weighted equally and 2) element-wise precision where each query is weighted equally globally. Following the official TAC evaluation, we use the second element-wise definition.

The primary evaluation measures for entity linking are macro-averaged accuracy (in the absence of NIL clustering) and B-Cubed+ F1.

$$P = \frac{\sum_{q \in Q} P_i}{|Q|}$$

$$R = \frac{\sum_{q \in Q} R_i}{|Q|}$$

Given this formulation of cluster precision and recall the official evaluation uses F1, as described previously.

3.4.3 Retrieval

Retrieval systems return a ranked list of results to users. There are evaluation measures for both binary relevance as well as graded relevance assessments. One simple evaluation

measure for binary relevance is *precision at k* ($P@k$), the precision up to the k-th result. For a single query it is defined as:

$$P@k = \frac{\sum_{i=1}^k rel_i}{|Q|}$$

where rel_i is the graded relevance of the document retrieved at rank i . For precision is a binary indicator, 0 if non-relevant and 1 if relevant. When computed over a set of topics, the mean $P@k$ is used. However, $P@k$ on its own does not capture the nuances of the ranked lists. It only examines the top k results. It is a set based measure that does not take into account the positions of the relevant documents in the ranking up to position k . Because of this, we also report results using *average precision*.

Average precision is a widely used retrieval evaluation measure. For a single query is defined as:

$$AP(q) = \frac{\sum_{i=1}^k P@i * rel_i}{|\mathcal{R}|}$$

where \mathcal{R} is the set of all relevant documents for the topic. Average precision takes the average of the $P@k$ values for each change in recall, where a relevant document is retrieved. It can be computed over the entire ranked list, but in practice it is computed up to a sufficiently large value. In this thesis we follow a widely used convention and evaluate it for the top 1,000 documents. When AP is used over a collection of topics it is referred to as *mean average precision*. It is defined as:

$$MAP = \frac{\sum_{q \in Q} AP(q)}{|Q|}$$

The TREC web corpora we use contain graded relevance assessments and focus on web retrieval. For these corpora, one of the widely used measures is *normalized discounted cumulative gain* at a given rank cutoff k ($nDCG@k$). It is defined relative to an optimal ranking.

$$nDCG@k = \frac{\sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i+2)}}{\mathcal{Z}_{\mathcal{R}}}$$

where $\mathcal{Z}_{\mathcal{R}}$ is the ideal gain computed from an ordering of the documents in decreasing order of relevance.

Another graded relevance metric used for web evaluation is *expected reciprocal rank* (ERR@k) proposed by Chapelle et al. (2009). ERR@k is inspired by the ‘cascade’ user model which assumes that users scan down a ranked list in order searching for a document which satisfies their information need. Expected reciprocal rank is the expected rank at which the user stops searching. It defines the probability of a user being satisfied with the i -th search result as:

$$P_i = \frac{2^{rel_i} - 1}{2^{rel_{max}}}$$

where rel_{max} is the highest relevance grade. Using this definition, expected reciprocal rank is computed as:

$$ERR@k = \sum_{i=1}^k \frac{P_i}{i} \prod_{j=1}^{i-1} (1 - P_j)$$

This measure was demonstrated to correlate better with user behavior in web search than nDCG@k and other metrics. The difference is particularly sharp for navigation and head queries.

In order to compare the output of two retrieval systems, we need to compare whether there is a statistically significant difference between them. To detect statistical significance for entity linking and document retrieval we use the Students’s paired t-test. The test makes the null hypothesis that the systems are identical. One known issue with this test is that it assumes a normal distribution of the system scores. It has been shown to be unstable for small numbers of queries (<50). The query sets we use in this thesis contain hundreds or thousands of queries. Smucker (2007) performs a detailed comparison of how this

significance testing method compares with others for the evaluation measures used in this work. If $\alpha < 0.05$ we conclude there is a significant difference between the retrieval systems.

CHAPTER 4

ENTITY-BASED ENRICHMENT

```
The traveller of steady head will delight to stand on
Pocumtuck Rock sheer above Eagle Brook Plain ...
out beyond the Bars, Indian Hole Squaw Hole, Bars Long
Hill, the Grindstone, and Sugar Loaf, spread out in
Nonotuck Valley the meadows of Old Hatfield and older Hadley
finally the brother peaks of Mount Tom and Mount Holyoke
stop the way, picturesque guardians of "Long tidal River,"
Quinetahacut.
```

Figure 4.1: Passage from Old Paths of the New England Border, pages 164,166.

In this chapter, we describe the framework for entity enrichment we use throughout this thesis. To illustrate the enrichment process, we use a real example of entity recognition, the task we study in Chapter 5. The local observation sequence, a sentence from a historical book, is shown in Figure 4.1.

4.1 Terminology

We first define the terminology we use throughout this chapter and in the thesis overall. A summary of the terms is given in Figure 4.2. First, we define an *observation*, which corresponds to a representation of text as a sequence of observed variables, $o \in O, o_1 o_2 \dots o_{|O|}$.

For each task we study in this work, and to generalize to other tasks, the definition of the observed variables in the sequence varies. When we discuss enrichment for a task we follow the naming convention used for observations in that community. For named entity recognition, the observed variables are words, or tokens, x each of which corresponds to

<i>observation, o</i>	An observed variable in text. This is an abstraction which may refer to a word, a mention, or a mention linked to an entity depending on the stage in annotation.
<i>mention, m</i>	A single occurrence of an entity in text, consisting of the name string and other possible attributes.
<i>entity, e</i>	A single thing or object in the world, which may occur as an entry in a knowledge base.
<i>linked mention, m</i>	An entity mention with additional attributes linking (disambiguating) it to one or more entities, which may exist as entries in a knowledge base.

Figure 4.2: Glossary of terms

a single observed variable, o . For the others in this thesis, the observations include *query mentions* for the task of entity linking and *query observations* for a search query in retrieval.

Based on local independence assumptions between variables (across sentences, documents, or queries) we refer to the observations in a particular sequence as ‘local’ observations. Observation and mentions that occur in other text sources are ‘non-local’. The ‘non-local’ observations may be within a document or query, but more importantly also includes observations from across text sources.

We focus on exploiting similarities between an observation and other similar entity *mentions* across large collections of text sources and a knowledge base. We define the topical similarity on an observation to other mentions with respect to a given query model and local context model created from the ‘local’ observations. .

4.2 Overview

Entity enrichment is a process that performs structured feature expansion of local observations from topically similar entity mentions in other documents and external knowledge resources. Entity enrichment can be performed at different levels of representation for any target observation. A target for enrichment could be a document, word, entity mention, user query, or other meaningful unit. The local observation, o may exist as part of a larger local context: O , which may be a sequence of observations, . Each observation o_i has an

associated vector of local features, f_{Loc} derived from o_i and the other observations in O . The result of enrichment is a new vector of non-local expansion features, f_{En} extracted from similar entity mentions in external sources.

The enrichment process consists of several important steps which vary for each task and collection. The steps in entity-based enrichment are: 1) enrichment triggering, 2) target model generation, 3) mention retrieval, 4) mention feature extraction, and 5) feature aggregation. In subsequent chapters, we instantiate this model for different tasks, with different characteristics for each. We now describe each of these phases in more detail.

4.3 Enrichment Triggering

Because entity-based enrichment identifies topically related entity mentions from external sources, there may be a significant cost to retrieving and analyzing the external data. As a result, we introduce an optional filtering step to perform enrichment only for a subset of observations. Enrichment triggering is the process of determining the observations for which enrichment should be performed for a particular task. In the simplest case each observation is enriched. However, for large sequences of observations, this may be infeasible and unnecessary. Instead of enriching every observation, one option is to identify difficult or ambiguous local observations for which enrichment would provide the greatest benefit.

To do this we define a binary decision function, g that determines whether to perform enrichment for an observation o_i .

$$g(o_i) = \begin{cases} 1 & \text{if } o_i \text{ is enriched} \\ 0 & \text{if enrichment is not performed} \end{cases}$$

The optimal decision function minimizes the overhead of enrichment and maximizes improvement in task effectiveness. The correct balance of these factors depends on the efficiency vs. effectiveness trade-offs of the application. Factors in designing the decision function may include the ambiguity of the observation as well as the cost of enrichment.

Enrichment is a form of external knowledge acquisition. The decision process for when and how to acquire external knowledge for information extraction tasks was studied in the context of resource-bounded information acquisition (RBIA) (KANANI and MCCALLUM 2012). In this framework, whether to perform resource-intensive actions is learned using reinforcement learning. Because this step of the enrichment process has received significant attention in other research, it is not our main focus. For this thesis we define simple task-specific heuristics and focus on the other aspects of enrichment.

4.4 Target Model Generation

Given an observation, o for which enrichment is triggered, we refer to the observation as an enrichment target, T . The first step in enriching the target observation is to extract a local context model. The goal of this phase is to identify informative features from the local context in O that are useful for identifying similar entity mentions.

The result of this phase is a model θ_T that contains features used for similarity comparison to identify relevant entity mentions. The components and construction of θ_T are highly dependent on the enrichment task and information collection. Although θ_T varies depending on the task, in this thesis it is often a multinomial distribution over words or entity mentions. θ_T may include positional information so that dependencies that incorporate phrases and proximity may be used.

```

<enrichment target> Hadley
<observations> The traveller...spread out in Nonotuck Valley the meadows of Old Hatfield
and older Hadley; finally the brother peaks of Mount Tom and
Mount Holyoke stop the way...Quinetahacut.
<word sequence> [The] [spread] [out] [in] [the] [meadows] [of]...[Quinetahacut]
<capitalized sequences> (The), (Pocumtuck Rock),... (Nonotuck Valley),... (Quinetahacut)

```

Figure 4.3: Example context model for the observed token ‘Hadley’ in entity recognition.

An example of local model generation for entity recognition is given in Figure 4.3. The example shows the model for the ‘Hadley’ observation token. Because the observation sequence in entity recognition is small, only a sentence, the model includes all of the

```

#combine(
  #sdm( Pocumtuck Rock )
  #sdm( Eagle Brook Plain )
  ...
  #sdm( Nonotuck Valley )
  #sdm( Old Hatfield )
  #sdm( Hadley )
  #sdm( Mount Tom )
  #sdm( Mount Holyoke )
  #sdm( Long )
  #sdm( River )
  #require( #all(hadley) )
)

```

Figure 4.4: Example mention retrieval query constructed from the ‘Hadley’ context model

observed words. It also includes sequences of capitalized words, which are likely to be other entities. Although not in the example, we observe that other features derived from the local observations could be included, include features from part-of-speech tagging or dependency parsing.

4.5 Mention Retrieval

The goal of mention retrieval is to identify text sources, D , containing entity mentions, \mathcal{M} that are similar to the enrichment target, T . In this step documents or knowledge base entries are ranked by their similarity to the enrichment target. To do this, the model of the target, θ_T , is used to generate a query, Q_T . One possible query generated from the model in Figure 4.3 is shown in Figure 4.4.

The target Galago query shown in Figure 4.4 has several important features. First, it uses the `#sdm` to indicate the sequential dependence model, which capture dependencies between adjacent words in meaningful semantic units, namely the other capitalized word sequences likely to be related entities. The local context is important because it focuses on the relationship of the target, “Hadley”, to other geographical features in the region. The

other words or features provide a topical focus of the context in which the target observation occurs. Second, the query contains a requirement that retrieved sources contain a matching observation, “Hadley”.

The query is used to retrieve and rank mention sources, $d \in D$, with similar mentions. For the entity recognition example, sentence retrieval is performed. The output of this step is a ranked list of text sources scored by the query model containing entity mentions from which enrichment features are extracted. A sample of the top ten similar sentences is shown in Figure 4.5. The next step is to extract features from the mentions in these sources.

4.6 Mention Feature Extraction

For each text source, $d \in D$ let \mathcal{M}_d be the related mentions of the target T_{md} . The set of all related mentions from the sources is $\mathcal{M} = \bigcup_{d \in D} \mathcal{M}_d$. The enrichment features are then extracted from the related entity mentions, $m \in \mathcal{M}_d$. Each task defines its own set of real-valued feature functions, $\psi(f_k, m)$, for each feature, $f_k \in \mathbf{f}$. In the simple case, $\psi(f_k, m)$ is a simple binary indicator function, for example to indicate the presence of a word or match in an external knowledge resource. This feature extraction phase differs from other related expansion models because features are extracted from individual entity mentions related to the enrichment target rather than words or features from the document overall. The type of relationship of the mentions to the target observation varies depending on the task, from exact-word matching, partial or complete name matching, or simply co-occurrence. As shown in the example query in Figure 4.4 for this task there is a requirement for the observations to share the same (normalized) string value.

For each identified entity mention, $m \in \mathcal{M}_d$ we perform feature extraction. The result of this process is a feature vector for each mention, \mathbf{f}_m . An example set of extracted features for the task of named entity recognition is shown in Figure 4.6. The extracted features can be from the entity mention itself or from its surrounding context, such as text or neighboring

entity mentions. The result of this step is the set of scored mentions, f_M , with extracted features $\langle m, p(d|Q_T), f_m \rangle$ for all mentions, M .

4.7 Feature Aggregation

Given the extracted features from all the related mentions, f_M with their extraction source, the output of this step is an aggregation into a single aggregated feature representation f_{En} . Aggregation of entity enrichment features is a two-step aggregation process. First, features are aggregated for each mention source, $d \in D$. This is important because it normalizes the contribution of individual sources so each source has equal contribution. Otherwise, long sources with many related mentions might contribute disproportionately to the feature weight. Next, the features for each source are combined, incorporating their similarity to the enrichment target, T . The aggregation model for enrichment varies depending on the type of features used for the task.

Many of the features used in this thesis represent feature counts. Consequently, we use an aggregation model similar to those in the relevance modeling (LAVRENKO and CROFT 2001) framework.

First, the features from similar mentions in the source document, M_d , are aggregated with respect to all the similar mentions in the source. An example of this is shown in Equation 4.1. Next, the feature values from individual sources are combined:

$$p(f_k|d) = \frac{\sum_{m \in M_d} \psi(f_k, m)}{|M_d|} \quad (4.1)$$

The per-source feature values are also aggregated. This aggregation incorporates the per-source value as well as the source similarity to the enrichment target as follows:

$$p(f_k|Q_T) = \sum_{d \in D} p(f_k|d)p(d|Q_T) \quad (4.2)$$

The result of this aggregated feature representation across all sources and their entity mentions, f_{En} . As in traditional relevance feedback techniques, a subset of the high probability features may be used for enrichment.

4.8 Summary

In this chapter, we introduced a framework for entity enrichment, which performs structured feature expansion of a local observation target. We outlined techniques for triggering enrichment, constructing a local model describing the target, generating a query model and retrieving similar entity mentions, extracting features, and finally aggregating features across sources. In the following chapters we instantiate this model for three different extraction and retrieval tasks.

North is the winding river, broad meadows and the villages of **Hadley**, Hatfield, Whately, with Sugar Loaf Mountain rising in the center of the valley.
 <score> P(D|Q) = 0.20

About the middle of September Captain Appleton with his company were marching from Deerfield to **Hadley**. In the neighborhood of Mount Sugar Loaf they stopped by a brook to pick wild grapes that hung temptingly on the vines about them.
 <score> P(D|Q) = 0.11

The broad meadows, and the village of **Hadley**, extending across the peninsular, from one bank of the Connecticut to the other, and the view down the Connecticut and across the meadows, here obtained, present to the beholder a scene of rural beauty rarely met with.
 <score> P(D|Q) = 0.07

But it was a sad state of things here in the valley, with Deerfield, Northfield, and Springfield destroyed and only **Hadley** and Northampton remaining.
 <score> P(D|Q) = 0.5

But on the east and south-east you have Amherst and Belchertown, and the Pelham Hills; on the south, the Holyoke and Tom Range; on the south-west, Northampton and Easthampton, and **Hadley**, with their rich surrounding meadows; on the west, close beneath you, lies quiet Hatfield, and a little farther north, the village of Whately; and behind them both rise the hills that make the eastern border of the Hoosac Range.
 <score> P(D|Q) = 0.03

There is good authority to believe, from records of the General Court in 1853, and other authentic sources, that the name we have chosen is the old Indian name of lladley, which then embraced what is now called **Hadley**, Amherst, Granby, South **Hadley**, Sunderland, Hatfield and Whately, and of course included this entire range of mountains.
 <score> P(D|Q) = 0.02

Another branch, coming from Mount Boreas, unites with that from Pelham and the stream passes obliquely across Amherst and **Hadley** and empties into the Connecticut at the foot of Mount Holyoke.
 <score> P(D|Q) = 0.02

From the Hotel are the drives to Mt Holyoke, Hockanum, Mt. Nonotuck, Mt. Tom, Sugar Loaf, Mt. Toby, Amherst, Easthampton, Holyoke City, South **Hadley**, and Florence, all places of interest in themselves.
 <score> P(D|Q) = 0.02

In the earlier part of the 17th century the Indians dwelling in Western Massachusetts near the Connecticut river belonged to four small tribes or clans, the Agawams, located at Springfield and West Springfield, the Waranokes at Westfield, the Xonotucks or Norwottucks at Northampton, **Hadley** and Hatfield, the Pocomtucks at Deerfield.
 <score> P(D|Q) = 0.01

To the north is **Hadley**, and beyond, Mt. Toby and Sugar Loaf, capped in the distance by the blue peaks of the Green Mountains.
 <score> P(D|Q) = 0.01

Figure 4.5: Example retrieved passages for the ‘Hadley’ target.

```

Mention source:
North is the winding river, broad meadows and the villages of Hadley, Hatfield, Whately,
with Sugar Loaf Mountain rising in the center of the valley.
<score> P(D|Q) = 0.20

 $\psi(f, m)$  Features:
prev_word2_villages, 1
prev_word1_of, 1
next_word1_hatfield, 1
next_word2_whately, 1
suffix_ey, 1
prefix_Ha, 1
token_category_LET-MIX, 1
prev_pos_IN, 1
cur_pos_NP, 1
next_pos_NP, 1
is_capitalized, 1
wiki_gaz_exact_location, 1
freebase_cat_city, 1

```

Figure 4.6: Subset of extracted entity recognition features f_m from related mention.

```

prev_word2_villages, 0.27
prev_word2_deerfield, 0.11
prev_word2_easthampton, 0.03
prev_word2_amherst, 0.02
prev_word2_city, 0.02
...
prev_word1_of, 0.27
prev_word1_to, 0.11
prev_word1_called, 0.02
...
is_capitalized, 1.0
token_category_LET-MIX, 0.82
token_category_LET-CAP, 0.13
...
prev_pos_IN, 0.40
cur_pos_NP, 0.90
next_pos_NP, 0.50
next_pos_NP, 0.30
wiki_gaz_exact_location, 0.90
freebase_cat_city, 0.85
...

```

Figure 4.7: Sample aggregated feature values recognition features f_{En} .

CHAPTER 5

NAMED ENTITY RECOGNITION

In this chapter, we describe entity enrichment for the task of named entity recognition. Named entity recognition is a pattern recognition task that assigns categorical entity labels (person, organization, location, miscellaneous) to a sequence of observed words. The goal is to infer a hidden label y from the observed token sequence x . The enrichment target in this task is the feature representation of each observed token, x_i .¹

5.1 Introduction

Despite the increased application of Natural Language Processing (NLP) on queries and documents to improve retrieval tasks, there is little work exploring the use of retrieval to improve extraction tasks. In this chapter, we use entity-based enrichment to improve the task of detecting and classifying named entities, commonly referred to as Named Entity Recognition (NER). Beyond NER, the enrichment model described in this chapter could also be used for similar sequence labeling tasks including part of speech tagging, syntactic chunking, and others. In these problems, we are given an input sequence of observed variables, \mathbf{x} , which consists of a sequence of words in a text document. For each observed variable, $x_i \in \mathbf{x}$ the goal is to infer a corresponding output label.

In most statistical sequence models the decision about the output label of a given token depends only on a small local window of adjacent text, typically a sentence. The local context that an entity occurs in may not provide enough evidence to accurately infer the

¹This chapter is partially based upon work published at the 20th ACM Conference on Information and Knowledge Management (CIKM '11) (DALTON *et al.* 2011).

output label. This problem is exacerbated by tables, lists, and other structures containing non-grammatical text with little or no contextual clues. The result of this is incorrect and inconsistent entity labeling. To improve effectiveness for these tokens, we investigate methods that leverage information from external or ‘non-local’ evidence, within and across documents.

To address this issue, we use the enrichment framework described in Chapter 4. To better estimate the features used to label a token. The enrichment framework used to expand the feature representation has several important properties that make it attractive for handling non-local dependencies in NLP tasks. First, in mention source retrieval, the local context of the token is used to find similar sentences. As we show in our retrieval evaluation, using local context is effective for ranking sentences, retrieving sentences with similar mentions that have matching labels, even for ambiguous tokens. Second, the number of non-local dependencies to similar mentions is controlled by varying the number of source sentences retrieved. Third, unlike existing models for non-local dependencies the features extracted from similar mentions are aggregated incorporating the similarity to the enrichment target. The result is more effective tagging. Finally, the model is efficient, because the number of non-local features from enrichment can be limited to only the most important or highest probability features.

The idea of tying labels and features across tokens has been explored in previous work modeling non-local dependencies, such the skip-chain CRF model (SUTTON and MCCALLUM 2004). However, efforts to model non-local dependencies directly in the graph structure result in complex graphical models with loopy graphs that require approximate inference methods, such as Loopy BP and Gibbs sampling. The use of approximate inference results in significantly slower performance (FINKEL *et al.* 2005). Consequently, these models are not often used in practice.

Another approach to incorporating non-local dependencies is based on copying and aggregating observed features (VILAIN *et al.* 2009; RATINOV and ROTH 2009). Copying

features allows the use of simple linear models where efficient exact inference techniques for training and decoding, such as the Viterbi algorithm, can be used. However, results using this approach in the past have been mixed. The results of Villain et al. (2009) show that feature copying improves the results on the CoNLL 2003 shared task, but not as much as they expect. One cause of errors that they highlight is ambiguous tokens that refer to the same entity but take on different labels depending on the context. For example, consider the word *China* which in: “China beat out Finland in the match...” is an ORG and “The Beijing Olympics took place in China.” where it is a LOC. Previous models treat all occurrences of a token identically without consideration of the context. Our enrichment approach addresses this problem by modeling the similarity of expansion sentences to the enrichment target.

Another problem with many existing models (SUTTON and MCCALLUM 2004) is that they only use non-local evidence within the same document. The result is that these models do not improve effectiveness on tokens that occur infrequently within a document. To address this problem the enrichment process we propose also utilizes features across documents. The enrichment process can leverage large collections of unlabeled text documents, such as the web, to improve effectiveness.

One of the stated design goals of NER systems is that they should be robust to unseen text. However, state-of-the-art systems perform poorly when evaluated on out of domain data. Liu et al. (2011) demonstrated that the effectiveness of the Stanford NER tagger trained on CoNLL data drops to 45.8% F1 when tagging entities from Twitter microblog documents. In our experiments, we find similar degradation in performance to 51% when evaluated on out-of-domain book data, namely the Deerfield collection of historical books. The behavior of these systems across on out of domain data results in a decrease in F1 score of approximately 40%. Our experiments show that models incorporating enriched feature representations are more robust than local models when evaluated on out of domain data.

The main contributions in this chapter are:

- describing an enrichment model for incorporating non-local dependencies;

- demonstrating that entity-based enrichment outperforms previous models of feature aggregation and consistently improves effectiveness;
- evaluating the effectiveness of varying mention source retrieval models to rank sentences based on the likelihood that shared tokens have the same entity label;
- showing that enrichment using external unlabeled data results in more significant improvement than using only labeled data; and
- demonstrating that models that utilize enriched features are more robust when evaluated on out of domain data, outperforming a leading sequence tagging system.

5.2 Non-Local Dependencies in NER

Sequence labeling tasks in natural language processing often make strong local independence assumptions. For example, many assume independence between observations across sentences. Extraction is performed on each sequence independently and in isolation. However, this results in incorrect and inconsistent labeling in information extraction tasks. In many cases, modeling the relationships between extractions can improve effectiveness.

Several recent efforts have focused on adding non-local dependencies, mostly within a single document, to penalize inconsistent labeling and enforce some degree of consistency. Finkel et al. (2005) show that predictions for the same entity are inconsistent within the same document and across the corpus. Sutton and McCallum (2004) use a skip-chain CRF with loopy BP inference to enforce consistent decoding among string-identical tokens in the same document. Finkel et al. (2005) penalizes inconsistent labeling within a document using Gibbs sampling. Bunescu and Mooney (2004) use a Relational Markov Network (RMN) to explicitly model long-distance dependencies and use loopy BP for inference. All of these techniques employ approximate inference techniques. Instead, of encoding the dependencies in the model, the enrichment model we propose avoids approximate inference, performing exact inference on copied features. An important difference in our model is

that instead of treating all dependencies equally, the enrichment framework incorporates the similarity between the target and non-local observations.

Another approach to global inference is two-pass or stacked architectures. A token which appears in an unindicative context in one sentence may appear in informative contexts in other sentences. In a two pass model the predictions of a first-pass system are used as global features in a second-pass model that “fixes up” the labeling (KRISHNAN and MANNING 2006; RATINOV and ROTH 2009). The simplest version of this approach enforces consistency in certain labelings by majority vote or other heuristics (MIKHEEV 1999). Other versions use nearest neighbor classification to incorporate predictions in other parts of the document or corpus (LIU *et al.* 2011).

Bendersky et al.(2010) tag sparse and ungrammatical web search queries using labels from top retrieved documents where instances are weighted using pseudo-relevance feedback. This is a two-pass model that incorporates similarity to the target sequence, an entire query. The structural annotation tasks they perform on queries include capitalization, part-of-speech tags, and segmentation. Our work has several important differences. First, although the enrichment framework we propose could be applied to detecting entities in queries, it is not the focus of this chapter. Detecting entities in queries has been studied by others (GUO *et al.* 2009). Similar to the enrichment model, the model incorporates evidence from retrieved documents and incorporates similarity to the enrichment target. However, instead of aggregating votes of label output, the enrichment method we propose copies low-level tagging features. As we discuss below, feature aggregation has been demonstrated to be more effective than vote aggregation for named entity recognition.

Two pass models fix mistakes, especially for frequent entities. The limitations of these models were recently examined by Villain et al. (2009). They fail when the first pass labels the instance incorrectly more often than correctly. Furthermore, for rare tokens the prediction information remains sparse and there may only be weak evidence in each sentence

considered in isolation. Their results find that copying low-level features is more effective than majority counts across several named entity recognition evaluation sets.

Our work on feature enrichment is similar to previous work on context aggregation, which copies features across token instances. Ratinov and Roth (2009) aggregate features for string-identical tokens within a fixed window size of 200, even across document boundaries. The idea of entity enrichment is most closely related to that of Villain et al. (2009), who copy “displaced features” across related tokens within the same document. Their method uses information gain to copy only the most predictive features for related tokens. It requires a pre-processing step over the entire corpus to identify these features over the corpus before training or decoding. The model suffers from ambiguous token contexts, introducing noise. In contrast, our enrichment framework incorporates sentence similarity to address issues of ambiguous contexts. Instead of information gain, our aggregation model based on highest probability features does not require a pass over the entire corpus, only a small subset of sentences.

5.3 NER Approach

The methods we propose can be incorporated into a variety of models used to infer output values in sequence labeling. For this work we incorporate our enrichment technique with a state sequence model based on Conditional Random Fields (CRFs) (LAFERTY *et al.* 2001). CRFs are a type of discriminatively trained undirected graphical model trained to maximize the conditional probability of output labels given an input observation sequence. Given an observed sequence of words x , the goal is to predict the values of the unobserved random variables y , which are the corresponding output labels. In this work, we utilize a linear-chain CRF with a first order Markov assumption made on hidden variables in the graph where only adjacent vertices are connected by edges. Just as with first-order HMMs, our model admits efficient inference using the forward-backward and Viterbi algorithms for training and decoding.

Feature
words = W_{i-2}, \dots, W_{i+2}
POS tags = o_{i-1}, o_i, o_{i+1}
W_i capitalization patterns
Character Prefixes = W_{i-1}, W_i, W_{i+1}
Character Suffixes = W_{i-1}, W_i, W_{i+1}

Table 5.1: Baseline NER features

CRFs are the state-of-the-art in many sequence modeling tasks (PINTO *et al.* 2003; LAFFERTY *et al.* 2001), and their effectiveness on NER tagging is competitive with the best reported by the LBJ NER tagger (KRISHNAN and MANNING 2006; VILAIN *et al.* 2009; RATINOV and ROTH 2009). Unlike generative models like HMMs, CRFs do not model the joint distribution $p(\mathbf{x}, \mathbf{y})$. Instead, they estimate $p(\mathbf{y} \mid \mathbf{x})$. The CRF framework allows the flexibility to integrate arbitrary features, including enrichment-based features. We train the CRF model using stochastic gradient descent (SGD). Our system is based on the open-source package LingPipe,². This baseline model corresponds roughly to the local Viterbi model described by Finkel *et al.* (2005). This class of models is widely used because of the models’ efficiency and simplicity.

The baseline local features used in the model include words within a window size of 4, adjacent word character prefixes and suffixes, part of speech tags, and capitalization patterns. The feature set is summarized in Table 5.1. For each of these features there is a binary feature function $f_k(x_i, \mathbf{x})$ that indicates the presence of the feature in the observed variables. For example, to indicate that the token is a noun, $f_{CUR_POS=noun}(x_i, \mathbf{x})$. We also evaluate models that incorporate external knowledge resources, such as Wikipedia gazetteers and Brown word clusters (BROWN *et al.* 1992). These features are used by a leading NER system, described by Ratino *et al.* (2009). In Section 5.5 we evaluate the these models with our proposed enrichment framework.

²<http://www.alias-i.com/lingpipe>

Feature	Description
notStop	Feature indicating absence of x_i in Lemur 418 stopword list
notBos	Feature indicating x_i is not the beginning of a sentence
isFirstCap	Indicator if the first character of x_i is capitalized
isCapOnly	Indicator if the first character of x_i is capitalized followed by lowercase (Aa+)

Table 5.2: Query trigger features

5.4 Entity-based Enrichment for NER

5.4.1 Enrichment Scope

Before exploring the different phases of enrichment for named entity recognition, we first define several enrichment scopes and relate this to previous work. We define the local observation sequence of tokens \mathbf{x} to be a single sentence. Likewise the enrichment sources for entity enrichment are sentences.

The source collection, C used for enrichment is an important factor in its effectiveness. It determines the scope of the non-local dependencies and the amount of information available. We now examine several corpus definitions and relate them to previous work.

Document The within document restriction defines the collection to be the sentences that occur in the same document as \mathbf{x} . In previous work this is the most commonly used model (SUTTON and MCCALLUM 2004; FINKEL *et al.* 2005; VILAIN *et al.* 2009). It is simple to implement because an entire document is typically available during labeling. Since documents that mention the same entity multiple times are likely referring to the same entity, there is strong evidence that the entity shares the same label. However, this definition does not consider dependencies between occurrences across documents. This hurts recall and is problematic for short documents and rare entities.

Fixed Token Window The fixed window definition restricts the retrieved sentences to ones that occur within a specified range of tokens in relation to the observed token, x_i . It can be used within documents, and has also been used for cross-document context aggregation by the LBJ NER tagger (RATINOV and ROTH 2009). The LBJ tagger uses a token window

size of 200. The fixed window cross-document collection definition is an ad-hoc heuristic developed based on the observation that documents close together in a newswire stream tend to be topically related. While effective, the heuristic is highly specific to the CoNLL data set and is unlikely to generalize to more general contexts. We include it in our experiments for completeness.

Global The global corpus definition utilizes all sentences in the source collection. The size and scope of the collection varies significantly. This definition supports dependencies across all mentions in every document. It is particularly useful when labeling infrequent entities.

External Beyond the source collection, other external sources of text are available. For example, large newswire collections, millions of scanned books, and the web. These sources vary widely in genre, topic, formality, and reliability.

5.4.2 Enrichment Triggering

Enrichment triggering is the process of determining the variables for which feature expansion should be performed. For entity recognition, enrichment for each observed variable x_i in \mathbf{x} is infeasible. Instead, we focus on enriching tokens that are likely to be named entities. As described earlier in the enrichment framework, we define a binary decision function, g that determines whether to generate a query, Q for each x_i in \mathbf{x} .

$$g(o_i) = \begin{cases} 1 & \text{if } x_i \text{ is enriched} \\ 0 & \text{if enrichment is not performed} \end{cases}$$

For our experiments we utilize several boolean combinations of the features in Table 5.2. For the data sets in these experiments the capitalization heuristics work well and have been successfully used in previous work (SUTTON and MCCALLUM 2004; FINKEL *et al.* 2005). Beyond capitalization, very common stopword tokens represent a large number of ambiguous observations, and queries generated from them are slower because they occur

in a large fraction of sources. In addition, tokens that are short (one or two characters) or all capitalized are likely abbreviations and are often ambiguous. Finding similar mentions for these highly ambiguous mentions may be difficult. The capitalization of tokens at the beginning of sentences are also ambiguous. To explore these options, we constructed several heuristic combination of these features that we evaluate in Section 5.5. We find that these simple heuristics are effective for our evaluation data sets, but other techniques may be needed for other types of text collections, such as microblog posts.

5.4.3 Target Model Generation

In this section we outline several methods for constructing a model of of an enrichment target, x_i from the local observations in \mathbf{x} . The goal is to generate a context model that is likely to retrieve sentences with similar mentions that contain the target variable x_i and share the same output label. Because at this stage in enrichment we have only text features, we focus on building a local textual model.

No context This model consists of only the current observed token, x_i . In previous work (RATINOV and ROTH 2009; SUTTON and MCCALLUM 2004; FINKEL *et al.* 2005) on modeling non-local dependencies, this is the only model utilized.

Adjacent tokens Beyond the observation itself, this model makes a first order Markov assumption and utilizes only adjacent tokens (x_{i-1}, x_i, x_{i+1}). We include this because it is an important feature used in NER classification.

All tokens All of the observed tokens in \mathbf{x} are utilized in the model. This utilizes the largest amount of local context information. However, it is also the most expensive to execute because these models may be quite large, resulting in slower execution.

Capitalized Tokens Although other identified entities in the sentence have not yet been detected, it is possible to observe a strong indicative feature, capitalization. In this context

model, we approximate using other co-occurring entities by using tokens that have the first character capitalized and the subsequent lowercase (Aa+). This captures likely entities and excludes ambiguous abbreviations.

5.4.4 Mention Source Retrieval

The goal of mention source retrieval is to identify text sources, $d \in C$, containing entity mentions, \mathcal{M} that are similar to the enrichment target. In the previous section, we examined the different elements of context used to model the target observation, x_i . From this model, we generate a query, Q_T , to retrieve sentences.

For retrieving similar sentences in NER, an important consideration is how the tokens are normalized. This includes case folding, stemming or lemmatization, and stopword removal. As we show in our experiments, in Section 5.5, features such as case sensitivity significantly impact the effectiveness of retrieving similar mentions.

5.4.4.1 Exact match

The simplest baseline model we evaluate is boolean set-based retrieval using exact string matching between the target observation and the source sequence. In this model, the query, Q consists only of the observed token, x_i . This model does not use any of the context in the surrounding sentence.

$$p(d|Q) = \begin{cases} 1 & \text{if } x_i \text{ is in } d \\ 0 & \text{if } x_i \text{ is not in } d \end{cases}$$

All sentences matching the query have the same uniform score. This is the method used in previous feature aggregation models (SUTTON and MCCALLUM 2004; VILAIN *et al.* 2009). For large collections the number of sentences retrieved can be very large, with thousands or millions of matches.

5.4.4.2 Unigram

The unigram model is equivalent to the Query Likelihood model that ranks documents according to the probability of relevance using a bag of words assumption of term independence. Using Dirichlet smoothing this is defined as:

$$\log p(Q|d) = \sum_{i=1}^{|Q|} \log \frac{f(q_i, d) + \mu \frac{c_{q,i}}{|C|}}{|d| + \mu} \quad (5.1)$$

where $f(q_i, d)$ is the frequency of the query term in the sentence, $c_{q,i}$ is the number of times a word occurs in a collection of documents, $|C|$ is the number of words in the collection, and μ is the smoothing parameter that is set empirically.

5.4.4.3 Term Dependence Models

To model dependencies between observations in the source model we generate Q using the sequential dependence variant of the Markov Random Field IR model (METZLER and CROFT 2005). It models dependencies between adjacent observations and includes phrases and word proximity. We described the SDM model in relation to log-linear models in Section 2.6.1. This model can be specified using the Indri³ query language as,

```
#weight( 0.8 #combine(United Arab Emirates)
0.15 #combine( #owl(United Arab)
#owl(Arab Emirates) )
0.05 #combine( #uw8(United Arab)
#uw8(Arab Emirates)) )
```

In this chapter the sequential dependence parameters are set according to those suggested by Metzler and Croft (2005) which were shown to be stable across collections.

5.4.5 Mention Feature Extraction

Given the retrieved source sentences selected by a technique in 5.4.4, we extract entity recognition features from these sources. We apply the same named entity recognition feature

³<http://www.lemurproject.org/indri>

extraction used previously for the local model on all of the target sentences. As discussed earlier in Section 5.3, each f_k is a binary feature function used for the features in the CRF.

5.4.6 Feature Aggregation

We use the two-step aggregation model outlined in Section 4.6. We observe that for the boolean exact match model all sentences have equal weight, so the enrichment features are simply an average of the feature values across the collection where the observation exists.

We utilize the non-local enriched feature distribution in addition to the original local features. We add the enriched features as new distinct feature functions in the model feature space. This allows the model to learn separate weights for local and enriched features. We then use the local inference methods for linear chain CRFs.

Adding features from enrichment approximately doubles the feature space used in the model. Over large collections the number of features can become prohibitively expensive. One technique to mitigate this issue is to only select a subset of the highest probability features for enrichment. Selecting a top-k most probable expansion features is commonly done in pseudo-relevance feedback and applies here as well. Similarly, another method to control the scope of the enrichment is to limit the number of feedback sources and use only the top-k sentences for enrichment.

5.5 Experiments

In this section, we report experimental results utilizing our enrichment model. First, we evaluate the mention source retrieval of various retrieval models. Second, we measure the effectiveness of the enrichment features on NER labeling in the CoNLL03 shared task. Third, we evaluate NER models that utilize external corpora as a source for enrichment features. Finally, we assess the robustness of the models by labeling an out of domain collection of scanned historical texts.

Trigger	# Queries	TP	FN	FP	TN	Prec.	Recall
isFirstCap	44906	33359	684	11547	158028	74.29	97.99
isFirstCap & notStop	41344	33273	768	8071	161504	80.48	97.74
isFirstCap & notStop & notBos	32552	27413	6628	5139	164438	84.21	80.53
CapOnly & notStop	33429	27912	6132	5517	164060	83.50	81.99
CapOnly & notStop & notBos	27240	24004	10040	3236	166341	88.12	70.51

Table 5.3: Query Trigger evaluation on the CoNLL training data. It compares boolean combinations of the features from Table 5.2.

5.5.1 Enrichment Triggering Evaluation

In the section we evaluate the effectiveness of several combinations of query trigger heuristics described in Section 5.4.2. Query triggering determines which observed variables are expanded. Ideally, it would occur when expansion improves labeling effectiveness; however, this is difficult to estimate directly. As a starting point, one heuristic we use is that feature enrichment should be performed if and only if the token is part of a named entity. This definition ensures that non-local entity information is considered in classification. The triggering evaluation results are shown in Table 5.3.

From the results, we observe that the heuristic utilizing capitalized letters has high recall. It captures all but 2% of entity tokens. The missing entity tokens are mostly stopwords that are part of a longer entity string (e.g. for in the sequence [Center for Intelligent Information Retrieval]), but has a significant number of false positives. The precision improves by removing stopwords, which are expensive queries to execute and are ambiguous tokens. The *CapOnly* heuristic excludes mixed and all-caps tokens which improves precision over *isFirstCap*. Although recall is reduced significantly, manual inspection shows that many of the missed tokens are abbreviations such as US, UN, and EU. *CapOnly* combined with excluding stopwords reduces the number of queries by 19%, reducing the number of false positives by half compared with the baseline *isFirstCap*. This is a significant savings in the number of queries executed. Most of the remaining false positives are temporal expressions such as month and days which are not labeled as named entities.

Retrieval	Zero Results	MAP	Mean Prec.	Relevant sentences	Returned sentences
Case Folding	2491	87.30	84.56	1,497,893	1,844,301
Case Sensitive	3112	90.57	88.35	1,161,370	1,297,002

Table 5.4: Evaluation of case normalization in retrieval using the Query Likelihood ranking and no context for the 33,429 queries.

Retrieval	MAP
CaseFold QL NoContext	87.30
CaseFold QL Adjacent	90.10
CaseFold QL All	91.14
CaseFold SD Capitalized	91.43
CaseFold SD All	91.70
CaseSens QL NoContext	90.57
CaseSens QL Adjacent	92.63
CaseSens QL All	93.50
CaseSens SD Capitalized	93.55
CaseSens SD All	93.92

Table 5.5: Evaluation of sentence retrieval using Mean Average Precision (in %). Various combinations of case sensitivity, retrieval model, and query generation method are evaluated. QL indicates Query Likelihood retrieval, SD indicates Sequential Dependence. The last word indicates the query generation method from Section 5.4.3.

The addition of the restriction to exclude tokens at the beginning of sentences, *notBos*, where virtually all tokens are capitalized, improved precision but resulted in a significant reduction in recall. Furthermore, capitalized tokens at the beginning of sentences are often ambiguous and enrichment can improve effectiveness by providing more features in less ambiguous contexts. We found tagging effectiveness improved by enriching these tokens.

We utilize the *CapOnly & notStop* combination for the remaining experiments. It is simple and provides a satisfactory trade-off between efficiency and recall.

5.5.2 Source Retrieval Effectiveness

For enrichment the effectiveness of the source retrieval phase is an important factor in enrichment effectiveness because the features are weighted by the model probabilities. We therefore evaluate various retrieval methods to determine which is the most effective. For

Approach	LOC	MISC	ORG	PER	ALL
Baseline	87.02	73.19	78.37	84.53	82.16
Stanford	86.11	77.78	78.50	85.39	82.62
Baseline + Brown	88.79	74.23	79.15	89.73	84.53
Stanford DistSim	89.64	77.35	81.08	90.63	85.88
Baseline + Brown + Wiki	89.59	74.48	81.49	91.91	86.08

Table 5.6: Phrase level F1 scores for base NER models described in Section 5.3 compared with the Stanford NER tagger on the CoNLL 2003 Named Entity Recognition test (b) set.

evaluation purposes a retrieved sentence, d , is defined to be relevant for a source query Q_{x_i} for variable x_i as follows:

$$Rel(d) = \begin{cases} 1 & \text{if } \exists x_i \in d \text{ s.t. } x_i = x_j \text{ and } y_i = y_j \\ 0 & \text{Otherwise} \end{cases}$$

where x_j and y_j are the corresponding variables contained in d . The above definition states that a sentence is relevant only if it contains a string-identical observed variable where the output labels have the same entity class.

The CoNLL newswire documents are indexed using the open-source Galago⁴ retrieval system. The documents are split into sentences using the boundaries provided and indexed to create a sentence level index. We perform stopping using the Lemur 418 stopword list and stemming using the Porter stemmer. Default Dirichlet smoothing was used with $\mu=2500$. For evaluation, the set of 33429 queries resulting from the query triggering method selected in Section 5.5.1 is used. The search index is loaded into memory for fast retrieval during tagging.

We first examine the impact of case folding on effectiveness. As previously discussed, capitalization is an important feature that strongly indicates a token is an entity. To utilize this we test case-folded and case sensitive retrieval. The results are shown in Table 5.4. As expected, case sensitive matching improves precision but decreases recall. The number of

⁴<http://www.galagosearch.org/>

relevant sentences retrieved decreases by approximately 20%. The number of queries with no results increases by 25%. It is notable that both models have very high MAP scores. The high MAP score indicates that most tokens in the CoNLL dataset are not ambiguous. Given the large number of sentences, the improvement in precision of case sensitive is more valuable than the decrease in recall.

Next, varying combinations of retrieval models described in Section 5.4.4 and context model generation in Section 5.4.3 are tested. The results of the evaluation on Mean Average Precision (MAP) are shown in in Table 5.5. The table shows that case sensitive retrieval results in consistent effectiveness improvements across all models. Using the entire sentence as context model performs the best. The target model using the capitalized words in the sentence performs only slightly worse than using all of the words in the sentence. This is significant because these queries are significantly more efficient to execute because they contain fewer terms that occur less frequently in the collection.

The best performing combination is the sequential dependence model using all words in the source sentence as the context model. As shown later in Section 5.5.3.3, this model also performs the best for NER feature expansion. This indicates that the relevance evaluation correlates with real NER improvements in the final combined system.

5.5.3 CoNLL NER Evaluation

In this section we measure the impact of adding non-local feature from enrichment to our baseline CRF model. We begin by evaluating the local baseline CRF models. Then, for comparison with previous work we evaluate features from exact match boolean retrieval. Finally, we evaluate effectiveness of enrichment models that incorporate more advanced mention source retrieval and ranking techniques.

5.5.3.1 Local NER Models

We now evaluate the baseline local tagging models. Table 5.6 shows NER systems and feature combinations using only local features on the CoNLL named entity recognition task.

Approach	F1	Error Red.
Local (baseline)	82.16	
FixedWindow	84.55*	13.4%
Global	83.86*	9.5%
Local (Brown + Wiki)	86.08	
FixedWindow	86.44*	2.6%
Global	86.11	0.2%

Table 5.7: F1 scores on CoNLL for feature enrichment using exact string matching for varying corpus scopes described in Section 5.4.3. The top is the baseline model with features from Table 5.1. The bottom results are for a stronger model with Brown clusters and Wikipedia features. Statistically significant over local models where indicated with a * with $p \leq .05$.

Approach	F1	Error Red
Local (Brown + Wiki)	86.08	
QL Capitalized	86.36*	2.0%
QL All	86.47*	2.8%
SD Capitalized	86.28	1.4%
SD All	86.60*	3.7%

Table 5.8: CoNLL F1 scores for feature enrichment using ranked source retrieval with the Global retrieval scope. (QL) indicates Query Likelihood and (SD) indicates Sequential Dependence retrieval models. The query context models used two variations: Capitalized includes only capitalized tokens, All has all tokens excluding stopwords. Significant differences over the local model with $p \leq .05$ are indicated with by *.

Approach	F1	% Err Red
Local (Baseline)	49.86	
Win200	51.41	3.1%
Global	55.36*	11.0%
Local (Brown+Wiki)	51.58	
Win200	51.06	-1.1%
Global	51.97	0.8%

Table 5.9: F1 scores of the NER model trained on CoNLL and evaluated on the Deerfield collection. The results show local systems and unweighted feature enrichment with varying collection scopes. The top is a tagger model with baseline features. The bottom is a stronger baseline model with word clustering and Wikipedia features. The differences are statistically significant with local models where indicated with a * with $p \leq .05$.

We compare the effectiveness of the our baseline tagger with the the Stanford NER system⁵. The base CRF model performance is comparable to the out-of-the-box Stanford system. Although these models are widely used for their efficiency, they are not state-of-the-art.

To the baseline system we add features from external knowledge sources. In particular, gazetteers from Wikipedia and Brown word cluster information. These resources are bundled with the freely available Illinois LBJ Named Entity tagger⁶. Consistent with the findings of Ratnov et al. (2009), the external knowledge resources provide significant improvement over the baseline model. These local NER models are the baselines we use to assess the impact of entity-based enrichment.

5.5.3.2 Exact Match Feature Enrichment

Next, we present the results of cross-document feature enrichment using sentences with string-identical tokens. Table 5.7 shows that enrichment provides consistent improvements over the local models. The FixedWindow expansion corresponds to the context aggregation method used by the LBJ tagger (RATINOV and ROTH 2009).

⁵<http://nlp.stanford.edu/software/CRF-NER.shtml>

⁶http://cogcomp.cs.illinois.edu/page/software_view/4

Approach	LOC	MISC	ORG	PER	ALL	% Error Red
ExactMatch	53.43	57.29	18.90	55.97	51.97	
LBJ (Win 200)	62.10	57.31	11.84	67.12	58.05	12.7%
QL All	64.62	53.47	23.64	63.79	59.31	15.3%
SD All	64.40	58.42	21.71	65.89	60.15*	17.0%

Table 5.10: F1 scores for CoNLL models evaluated on the Deerfield collection. The table compares global ranked feature enrichment models compared with baseline exact string matching. We compare against the state-of-the-art LBJ NER model that uses Fixed Window feature aggregation. All differences are statistically significant over the baseline ExactMatch model with a with $p \leq .05$, a * indicates significance over LBJ.

For the baseline retrieval system, the FixedWindow expansion method provides a 13.4% reduction in F1 error on the CoNLL dataset. The global enrichment model using all sentences in the collection provides a smaller 9.5% reduction. FixedWindow outperforms unweighted global feature enrichment. FixedWindow restricts the sentences to match those near the source sentence in the news stream. It exploits temporal (and therefore topical) locality in the CoNLL dataset. It does not perform well on collections that do not have this property, as we show later in the Deerfield evaluation. Neither aggregation method applied to the baseline model outperforms a stronger local model that uses Brown word clustering and Wikipedia gazetteers.

The results of adding exact match enrichment features to a stronger model incorporating Brown and Wikipedia is shown in the bottom of Table 5.7 there is a small, but significant improvement using the FixedWindow model. The expansion with the global retrieval over all sentences provides no significant benefit. The unweighted global aggregation has less topical cohesion and the unweighted enrichment contains more noise. The exact match model acts as a global prior for a token. This can be problematic for ambiguous tokens. We now explore the use of enrichment models using stronger source retrieval models, those that utilize local sentence context to address the problem of ambiguity.

Approach	F1	Error Red.
CoNLL SD AllTokens	86.60	
CoNLL QL AllTokens + Ext100	86.66	0.4%
CoNLL SD AllTokens + Ext50	87.01*	3.1%
Deerfield SD AllTokens	60.15	
Deerfield QL AllTokens + Ext100	60.07	-0.2%
Deerfield SD AllTokens + Ext50	61.22*	2.7%

Table 5.11: F1 scores for external feature enrichment including a 50k document subset of the RCV1 reuters news collection. Ext100 indicates 100 feedback sentences, Ext50 indicates 50 sentences. A * indicates significance over non-external model with $p \leq .05$.

Approach	F1 Score	Error Red.
CoNLL Best Local	86.08	
CoNLL Expansion	86.60	3.7%
CoNLL Expansion + External	87.01	6.7%
Deerfield Best Local	51.58	
Deerfield Expansion	60.15	17.7%
Deerfield Expansion + External	61.22	19.9%

Table 5.12: Summary Table comparing the F1 score of the strongest models in each category, a purely local model incorporating word clustering and gazetteers, a model using ranked feature enrichment models, and enrichment including an external corpus. All results are statistically significant with $p \leq .05$.

5.5.3.3 Ranked Feature Enrichment

The results for feature enrichment from ranked retrieval are shown in Table 5.8. Because the corpus is small all sentences are used for enrichment. Unlike the exact match based enrichment, the results show that ranked enrichment models result in significant improvement over the strongest local NER model. The SD AllTok combination provides a 3.7% reduction in error over the best performing local model.

The models with the AllTok context outperform models using only capitalized tokens. The Sequential Dependence model provides a small improvement over Query Likelihood. The models using AllTokens outperform the exact match model with the 200 token fixed window described in the previous section.

5.5.4 Deerfield Evaluation

In this section we evaluate the robustness of the models trained on newswire by evaluating them on the collection of scanned books described in Section 3.1.2. For the Global retrieval scope all the sentences in the 20 books are indexed. Sentence splitting is performed using the OpenNLP MaxEnt classifier.

The results for the evaluation on the Deerfield dataset are shown in Table 5.9. The results show that the F1 score of the tagger drops by approximately 40% compared with the CoNLL results. We investigated the errors and found that many of errors are due to sparsity in the book domain. A significant number of the entities in the book collection are not present in the newswire training collection. Our error analysis finds location entities are often confused for people. Investigating the feature values for these errors, we found that for unseen tokens the tagger relies heavily on the class prior, which is biased towards person labels in the newswire data. We now show the impact of feature enrichment on addressing these problems.

5.5.4.1 Exact Match Enrichment

Table 5.9 shows that expansion using Fixed Window of 200 tokens does not improve effectiveness significantly. The Global scope outperforms the Fixed Window method when applied to the baseline model.

It is curious that global enrichment model does not significantly improve the stronger local model that incorporates Wikipedia based gazetteers. In fact, the model performs worse than enrichment with to a weaker recognition model. We believe this is due to the phenomena of model undertraining (SUTTON *et al.* 2006) where the strong Wikipedia features in the newswire domain result in the model underweighting for token and context features.

5.5.4.2 Ranked Feature Enrichment

The results for ranked feature enrichment models are shown in Table 5.10. The enrichment models using ranked source retrieval result in very substantial improvements in NER effectiveness. The Sequential Dependence model using a query generated from the entire sentence results in a 17% reduction in error. It outperforms the LBJ Layer 1 model which is currently the best performing NER tagger on newswire data. The evaluation indicates that feature enrichment results in a model that is more robust across domains than local models.

The improvement in model effectiveness from enrichment does not address OCR errors. We only copy features for identical observed tokens. Relaxing this constraint to copy features for similar strings could potentially improve accuracy further for these tokens, which is an area for future work.

5.5.5 Enrichment using External Collections

Enrichment can also be used to improve NER effectiveness using unlabeled data from external collections. The previous experiments utilize small text collections used in NER evaluation. The labeled CoNLL data contains less than 20 thousand sentences. We now explore enrichment models that use features from larger external collections of text.

5.5.5.1 Reuters RCV1 subset

As an external source for feature enrichment we use a subset of the Reuters RCV1 collection (LEWIS *et al.* 2004). RCV1 consists of Reuters newswire data collected in 1996 and 1997. It contains documents from the same source and time period as the CoNLL data set. We use the first 50,000 documents of the collection. The RCV1 subset contains 931,822 sentences and 20.5 million words.

5.5.5.2 Evaluation

In previous experiments all of the sentences in the collection were used without a retrieval cutoff because of their limited size. For these experiments, enrichment is performed using only a subset of the top ranked sentences. We experiment with the number of retrieved sentences and report results using the top 50 and 100 sentences.

The results on both the CoNLL and Deerfield collections are shown in Table 5.11. The results compare against the best performing feature expansion models that does not utilize external data. The Sequential Dependence model with 50 feedback documents results in significant improvement in both the CoNLL and Deerfield evaluations. It provides a 3.1% error reduction in CoNLL and a 2.7% error reduction in Deerfield.

The model using 100 feedback documents and QL retrieval does not significantly improve effectiveness and slightly hurts effectiveness on the Deerfield data. We are unsure why this model does not perform as well, especially on the CoNLL data. It is possible that the larger collections contain more ambiguous tokens. Also, the larger number of feedback documents may introduce noisy features from off-topic sentences. More error analysis is needed to understand this behavior. We note that enrichment with the QL retrieval model is less effective than the Sequential Dependence model. For the Deerfield data, the additional newswire data may not contain the topics in the dataset and therefore may not be as useful for expansion.

Despite mixed results, the external feature expansion model results in the overall best performing system.

5.6 Summary

In this chapter we applied the enrichment framework to the task of named entity recognition (NER). The enrichment framework induces long-range cross-document dependencies between similar mentions using retrieval. It uses weighted feature copying from topically similar passages. In addition to showing that enrichment achieves statistically significant

improvements on in-domain accuracy, we show it results in a more robust entity detection model, significantly surpassing other methods when evaluated out of-domain data (Contribution 2).

The enrichment framework allows us to leverage large external sources of unlabeled data. The results show a 6.8% error reduction on newswire and a 19.9% error reduction on out-of-domain book data for named entity recognition. A summary of the results is presented in Table 5.12. In addition to showing that enrichment can achieve statistically significant improvements on in-domain accuracy, we show it results in a more robust model, significantly surpassing other context aggregation methods when evaluated out of-domain data.

In this chapter we use enrichment for detecting named entities. We used the enrichment framework to improve the effectiveness of detection and classification using cross-document evidence. The result is a text document annotated with mentions, M , of entities. In the following chapter, we apply the enrichment framework to the task of disambiguating the entity mentions to external knowledge sources, like Freebase and Wikipedia.

CHAPTER 6

ENTITY LINKING

```
When Chris Foy's final whistle reverberated around Villa Park in
May and Newcastle United were relegated to the Coca-Cola
Championship, Toon fans would have been forgiven for fearing
their club was beginning to fade into footballing obscurity.
```

Figure 6.1: Excerpt from TAC document with linking query for [Toon] entity.

In this chapter, we describe entity-based enrichment for entity linking. Entity Linking is the task of mapping an entity mention (name) in a document to entities in a knowledge base. In the previous chapter, we focused on detecting and classifying named entity mentions. In this chapter, we focus on linking these detected mentions to an external knowledge base, such as Wikipedia or Freebase. Although much of the work in this area has been done in the NLP community, the task can be viewed as a form of entity retrieval, where each mention is a query with one relevant document. One of the key differences between entity linking and retrieval is that the entity mentions occur in the context of a document, with rich contextual evidence that can be used for disambiguation and generating a contextual entity model.

One of the crucial problems for entity linking is identifying relevant disambiguating context in the mention document. The disambiguating context includes the words, but more importantly, other entity mentions (and their associated entity links). However, not all entity mentions are equally helpful for disambiguation. To address this issue we introduce the neighborhood relevance model which uses entity-based enrichment to identify the salience of entity context from similar mentions in other documents. We show that the enrichment-based model is more effective than local document context alone for ranking KB entities.

Experiments on the TAC KBP entity linking task demonstrate that when incorporated into the full entity linking system, it is one of the best performing systems for mentions that are linkable to the knowledge base.¹

In the last section of this chapter we study other types of enrichment useful for entity linking. An important area of study in entity linking task for the TAC KBP linking task in 2013 was linking informal forum documents. We study the effectiveness of enrichment from informal external resources, Urban Dictionary and metadata from other linked entities.

6.1 Introduction

Entity linking is important because most content created is unstructured text in the form of news, blogs, forums, and microblogs such as Twitter and Facebook. A key challenge is to link these unstructured text documents to the Web of Data. Entity linking bridges the structure gap between text documents and linked data by identifying mentions of entities in free text and linking them to knowledge bases. It enriches unstructured documents with links to people, places, and concepts in the world. Entity linking is a fundamental building block that supports a wide variety of information extraction, document summarization, and data mining tasks. For example, linked entities in documents can be used to expand existing knowledge base entries with new facts and relationships.

The major challenge in entity linking is ambiguity. An entity mention in text may be ambiguous for a wide variety of reasons: multiple entities share the same name (e.g., Michael Jordan), entities are referred to incompletely (e.g., Justin for Justin Bieber), by pseudonyms or nicknames (Christopher George Latore Wallace is also known as The Notorious B.I.G.), and are often abbreviated (e.g. UW for the University of Wisconsin as well as University of Washington).

¹This chapter is partially based upon work published at the 10th Conference on Open Research Areas in Information Retrieval (OAIR '13) (DALTON and DIETZ 2013a).

The entity linking problem has been studied over several years in the TAC Knowledge Base Population venue with the following task definition:

Entity Linking: Given a string mention m_q in a document, predict the entity e in the knowledge base which the string represents, or NIL if no such entity is available.

A typical entity linking process has four phases: 1) query expansion, 2) candidate generation, 3) entity ranking, and 4) handling NIL cases. The goal of the first two steps is to achieve a high-recall set of entities. Given the candidate set, most effective approaches (LEHMANN *et al.* 2010; CUCERZAN 2011; RATINOV *et al.* 2011) leverage contextual evidence, including neighboring entities, as disambiguating evidence in step 3. One issue is that the candidate generation step is often performed using string matching heuristics, resulting in large candidate sets that may contain hundreds or thousands of entities for ambiguous matches. The connection between candidate generation and ranking are often separated and not well aligned.

We advocate an information retrieval approach that uses one probabilistic model for steps 1-3. We introduce our linking system, KB Bridge. Supplementary materials for this work is available on the KB Bridge website². Existing entity linking methods only employ IR to a minor degree. We model the entire linking task as a retrieval problem. The graphical modeling framework allows us to ground our work on models from both information extraction and information retrieval.

For a given entity mention, the correct knowledge base entry is likely to share important pieces of contextual information: lexically similar names, shared topical similarity reflected in word usage, and similar patterns of relationships to other entities.

Entity linking provides some unusual challenges. Document retrieval is often performed with short keyword queries, with little or no local context information. In entity linking, the query is an entity string embedded in a longer document, providing an abundance of context

²<http://ciir.cs.umass.edu/~jdalton/kbbridge>

which could be leveraged. However, not all context is equally helpful, either because of vocabulary mismatch, ambiguity, heterogeneity in topic, or random cooccurrences. Consider the example “ABC shot the TV drama Lost in Australia.” with the task of linking “ABC” to the entity “American Broadcasting Company”. The named entity span “Australia” is not useful for disambiguating ABC, and would likely misdirect the model towards an incorrect linking to the “Australian Broadcasting Corporation”.

To address this problem, we introduce an entity-based enrichment framework, which we refer to in this chapter as the *neighborhood relevance model*. We use it to estimate the salience of contextual entities with the goal of re-weighting entities in the target mention’s context model. The model uses the entity-based enrichment model we describe in Chapter 4 to identify evidence from similar entity mentions in other documents.

The main contributions described in this chapter are:

- An unsupervised model for entity linking based upon entity retrieval that provides competitive performance out-of-the-box.
- A unified retrieval based approach to linking combining candidate generation and ranking in a single retrieval framework, with more than 95% recall in the highest ranked 10 entities.
- A mention-specific enrichment model for identifying salient contextual evidence from across-document evidence.
- Empirical evaluation of the entity-based enrichment model in combination with a supervised learning to rank framework on the TAC KBP Entity Linking task, resulting in one of the best overall system effectiveness for entities linkable to the knowledge base.

6.2 Related Work

Early work on entity linking was performed by Bunescu and Pasca (2006) and Cucerzan (2007) to link mentions of topics to their Wikipedia pages. In contrast to their models, we focus on a retrieval approach that leverages text based ranking without exploiting extensive Wikipedia-specific structure.

Our work is related to that of Gottipati and Jiang (2011) who take a language modeling approach to entity linking. They expand the original query mention with contextual information from the language model of the document. We use the local weighting as a starting point for estimating the entity salience and compare against it as a baseline.

It is also related to previous work on document expansion in speech retrieval, which uses relevance feedback to expand document models (SINGHAL and PEREIRA 1999) with evidence from similar documents. They find that there are two main effects from expansion: 1) reweighting terms that exist in the document and 2) adding new terms. They found that the majority of the improvements came from reweighting existing terms. Terms with equal counts in the document receive equal weight, but after expansion these are reweighted based upon the presence of those terms in related documents. Consequently, in our work we focus our efforts on reweighting. Our work differs in several important ways. First, instead of reweighting terms from similar documents the model that we propose focuses on weighting associations between a particular entity mention and other entity mentions in the document. And second, instead of performing expansion with the nearest neighbor documents, we perform enrichment focused on similar entity mentions.

Entity linking has been studied in a variety of recent venues. At INEX the “Link the Wiki” task explored automatically discovering links that should be created in a Wikipedia article (HUANG *et al.* 2008). More recently, it is one of the principle tasks studied at the ongoing Text Analysis Conference Knowledge Base Population track (TAC KBP). Ji *et al.* (JI *et al.* 2011; JI and GRISHMAN 2011) provide an overview of the recent systems and approaches.

Instead of linking individual mentions one at a time, recent work (CUCERZAN 2011; RATINOV *et al.* 2011; STOYANOV *et al.* 2012; KULKARNI *et al.* 2009; HOFFART *et al.* 2011) focuses on linking the set of mentions, M , that occur in the query document d . These models perform joint inference over the link assignments to identify a coherent assignment of KB entries. In our work we leverage the set of mentions M_d in the document as context in an information retrieval model. In this work we instead focus on identifying salient entity mentions in the context, because mentions in the document may be spurious or only tangentially related. This is especially true if the document contains multiple topics.

6.3 Mention Context Model

One of the key defining differences in entity linking is that the query entity mention is embedded in a document. In this section, we describe the contextual model of the entity that we extract from the document to represent the target query entity. The contextual model we use in this work has four main components: the query mention, name variations, surrounding sentence words, and ‘neighbor’ entity mentions.

The first and universally used piece of information is the entity mention string itself, m . The entity mention, m , consists of a string, (e.g. [Toon]) and a generic entity type (Organization). The problem is that in isolation the query mention may be highly ambiguous. In fact, the TAC KBP organizers specifically focus on these mentions because they are difficult. To address this issue, most entity linking systems expand the mention representation using evidence from the containing documents. This is a form of entity-specific query expansion.

The next piece of context that is widely used is entity name variations, V . In this case, the goal is to identify other aliases of the entity that occur in the document, which may be less ambiguous. One way this is done is through within-document coreference resolution. In this work, we use simple string matching heuristics. The first heuristic identifies variations based on other mentions containing the target based on overlap, e.g. [Toon] would be similar

to [Toon Network]. The other main heuristic handles abbreviations, for example [ABC] would match [Amherst Brewing Company] by looking for matching sequences of capitalized words (ignoring stop words) in other entities.

Another commonly used piece of context is the text in the document. Because the documents are long and may contain topics that do not directly relate to the target entity, we focus on the local context around the entity mention. Similar to the previous chapter on recognition, we use all of the words in the surrounding sentence containing the mention. We refer to this sentence context as S .

One of the most successful models, used widely by leading entity linking systems is modeling the representation of the query mention using other related entities, M in the document (MILNE and WITTEN 2008; LEHMANN *et al.* 2010; CUCERZAN 2007; RATINOV *et al.* 2011; MCNAMEE *et al.* 2012). Like the query mention, these mentions contain both the entity names and their types from mention detection. We refer to these mentions as entity ‘neighbors’.

Similar to NER, entity linking systems utilize the neighboring mentions as part of a two-pass linking model. In the first pass linking is performed for each mention in isolation. Then in the second pass features from disambiguated mentions (or their entity candidate distribution) are used as features. This allows the linking models to use features of the entities as part of the document representation, including the entities’ associations (via links) and category or type information. The goal is to select a coherent group of entities for the entire document. This is sometimes referred to as ‘joint assignment’ or ‘collective entity resolution’ because groups of entities are linked together to maximize coherence. For joint assignment models, the idea is that knowledge base entities which are mentioned in the same document are also likely to be structurally related in the knowledge base.

In this chapter, we use a simple and effective single pass model that incorporates a representation of the neighboring mentions using their names. Given these neighboring entity mentions, using them effectively for disambiguation is a key challenge. As we

show later in our experiments, uniformly weighting mention strings is not effective. To address this, we focus on weighting the association of the neighboring entities with the target mention. We refer to this as a neighbor’s salience, ρ . The $\rho_q(m)$ ranges on a scale between 0 and 1. If the salience $\rho_q(m)$ is 0, we want to remove the effect of $\phi^{\text{me}'}(m, e)$ on the likelihood function.

6.3.1 Local Neighborhood Model

One approach used in other work is to treat each entity in the neighborhood uniformly. These treats each entity (a group of coreferent mentions) equally. One proposed work goes beyond this and uses the mention frequency in the document for weighting. Gottipati and Jiang (2011) build a multinomial language model of entity mentions from the local document d with occurrence count $n_{m,d}$. We refer to this simple estimation technique as the local model. In this model, entities that are mentioned more often have higher weight.

$$\rho_q^{\text{local}}(m) = \frac{n_{m,d}}{\sum_{m'} n_{m',d}} \quad (6.1)$$

They also evaluate weighting schemes that incorporate distance to the target mention, but found that these did not significantly improve the results over using the entire document. However, because the local document entity evidence may be sparse or multi-topic, the model may be sub-optimal. For example, if the target entity is not the main focus of the document then co-occurring entities are not relevant for disambiguation and may actually lead to worse performance. In contrast, the enrichment model that we propose in this chapter uses the similarity of the mention contexts as a means of identifying related entity mentions.

```

<enrichment target> Toon
<name variations>
<sentence words> [When] [Chris] [Foy’s] [final] [whistle] ... [footballing] [obscurity]
<neighbor mentions> [Chris Foy] [Villa Park] ... [Coca-Cola Championship]

```

Figure 6.2: Example context model for [Toon] entity target from Figure 6.1

6.4 Entity Linking Retrieval Model

In this section we describe the ranking model we use for entity linking. We formalize the model using graphical models for entity linking as developed in the information extraction community combined with graphical models used for information retrieval. We utilize the fact that query models, such as query likelihood and the sequential dependence model (METZLER and CROFT 2005), have an underlying graphical model that gives rise to a score of a document.

For entity ranking, we use models that fall into the log-linear framework described in Section 2.6.1. Many complex factor functions are possible, but for the remainder of this publication we use two simple and effective factor functions ϕ : Factor $\phi^{\text{me}}(m, e)$ encodes matches of the mention m and terms from the surrounding text in any field of the Wikipedia article of the candidate entity, e , this includes names as well as the full-text of the article. Factor $\phi^{\text{me}'}(m', e)$ matches the string representation of m' in the full-text and titles entities that link to (and are linked from) the Wikipedia entry of e .

The factor ϕ^{me} formalizes our intuition on compatibility between the query mention q and the true entity e . This includes name matches of the string representation of m with names listed in the knowledge base (e.g. title, redirect, anchor text). We further extend it to other similarity measures that are independent of the neighbor mentions (which are represented by $\phi^{\text{me}'}$).

Name variations v of the query string can be extracted from the source document. This is especially important if the query string is an acronym or an ambiguous reference to the entity. We also incorporate the surrounding words, s , of the sentence containing the query mention or one of its name variations.

We introduce separate weight parameters λ^{Q} , λ^{V} , λ^{S} to individually control influence of name-matches of the query mention, name-matches of name variants v , and sentence context respectively. Accordingly, we model the factor function $\phi^{\text{me}}(q, e)$ by the likelihood

of a graphical model itself, represented by a log-linear function of potential functions ϕ^{name} for name-similarity and ϕ^{sent} for sentence context.

The resulting model for ranking the candidate entities for the target query mention consists of the the mention string, q , the name variations, V , the contextual words, S , and the neighboring entity mentions, M . The equation for this model is given in Equation 6.2.

$$\begin{aligned} \log \mathcal{L}(e) &= \lambda^{\mathbf{Q}} \log \phi^{\text{name}}(q, e) & (6.2) \\ &+ \lambda^{\mathbf{V}} \frac{1}{V} \sum_v \log \phi^{\text{name}}(v, e) \\ &+ \lambda^{\mathbf{S}} \frac{1}{S} \sum_s \log \phi^{\text{sent}}(s, e) \\ &+ \lambda^{\mathbf{M}} \frac{1}{M} \sum_m \left(\rho_q(m) \log \phi^{\text{me}'}(m, e) \right) \end{aligned}$$

Using log-linear models for factors ϕ with features that are readily available in the Indri and Galago³ query languages. Specifically, we use the sequential dependence model (METZLER and CROFT 2005), which is a query model for modeling dependencies between adjacent query words.

We use the sequential dependence model for matching the different query elements to the target entity. The matching includes several different elements. The name-match factor $\phi^{\text{name}}(q, e)$ tests all of the entity's indexed document representation for the presence of the string representation of the query mention. We do the same for the words, ϕ^{sent} and neighboring contextual entities $\phi^{\text{me}'}$. With these feature functions, the full ranking model is given in Equation 6.2. An example Galago query is given in Figure 6.3.

³<http://www.lemurproject.org/galago.php>

```

#combine:0= $\lambda$ Q:1= $\lambda$ V:2= $\lambda$ S:3= $\lambda$ M(
  #sdm( Toon)
  #combine()
  #combine(final whistle reverberated relegated
    fans forgiven fearing club beginning football obscurity)
  #combine:0 =  $\rho(m_0) : \dots k = \rho(m_k)$ (
    #sdm(chris foy), ..., #sdm(coca cola championship)
  )
)

```

Figure 6.3: Contextual query for [Toon], occurring in the sentence from Figure 6.1

6.5 Entity-based Enrichment for Linking

We now discuss methods for estimating these salience weights $\rho_q(m')$ in an unsupervised manner. The idea is to assume a high salience of a neighbor m' for the query mention m_q , if both are frequently mentioned together. It is important to note that even unambiguous mentions are not necessarily useful for disambiguating other mentions.

The idea is that a neighbor is important if it occurs frequently in the context of the query mention within the document as well as across other documents that are topically related and contain mentions of the query mention q .

Query Triggering

Like named entity recognition, a document may have tens or hundreds of entity mentions. Performing enrichment for every mention is costly. The query mentions in the TAC dataset are selected with a bias towards difficult and ambiguous mentions. Consequently, we perform enrichment for all query mentions. We do not evaluate triggering for this task. In practice, many linking models are two-pass systems, where an initial assignment of mentions to entities is performed for all mentions. Based upon the initial linking, features such as

ambiguity and link probability can be estimated and used to determine whether enrichment is required.

Target Model Generation

The model for generating the local mention context consists of the same components used to link entities to the knowledge base. The context generation process is described above in Section 6.3. As an initial estimate for the neighbor importance, the salience is weighted using the local document model, $\rho_q^{\text{local}}(m)$.

Mention Source Retrieval

The goal of mention source retrieval is to identify text sources, D , containing entity mentions, \mathcal{M} that are similar to the enrichment target. In this case, we perform retrieval over the source corpus' documents. Similarly, the source retrieval models we use for ranking entities in the knowledge base described above in Section 6.4 we also apply to ranking mention sources.

The reason we retrieve documents for mentions sources is because it was not possible to perform NLP processing on the entire corpus. Instead, we identify documents with a high likelihood of containing relevant mentions. We then perform entity mention detection only on this focused subset of documents.

Mention Feature Extraction

Given the ranked documents and the detected entity mention, we extract features from each of these sources, $d' \in D$. We first identify similar mentions of entities that are likely to be coreferent with our target mention. Similar mentions are found by matching their name surface forms to the target query mention and its name variations, $(q$ and $v)$. For each of these similar mentions, we extract co-occurrence counts of entities in their neighborhood, n_m, d' . Although other features may be useful, the primary feature we focus on in this work is estimating the salience of the neighbor entities to the query mention, $\rho_q(m')$.

Feature Aggregation

We use the two-step aggregation model described in our framework of Chapter 4. The first per-source aggregation follows Equation 4.1, which corresponds to the local model given in Equation 6.1. The second aggregation is a weighted combination of these models incorporating the similarity to the target mention, as in Equation 4.2.

We use the retrieval probability of the document as an estimate of how similar the mentions are to the target query mention. As most retrieval frameworks return only unnormalized (rank-equivalent) retrieval scores $\mathcal{L}(d)$, the estimate has to be approximated with $\frac{\mathcal{L}(d)}{\sum_{d' \in D} \mathcal{L}(d')}$.

Counting the occurrence frequency $n_{m,d}$ of string-identical mentions of m , we build a multinomial language model across the pseudo-relevant documents with relevance-model weighting as follows.

$$\rho_q^{\text{nrm}}(m) = \frac{1}{\sum_{d' \in D} \mathcal{L}(d')} \sum_{d \in D} \frac{n_{m,d}}{\sum_{m'} n_{m',d}} \mathcal{L}(d) \quad (6.3)$$

The result is a multinomial language model over the neighbor mentions.

6.6 KB Bridge: Entity Linking System

In this section we describe KB Bridge, our retrieval-based entity linking system which is implemented using the Galago search engine and the MRF information retrieval framework. The system links entity mentions in the source document to knowledge base entities. The ranking of the entities is a two-stage process. First, entities are ranked using the Galago retrieval model described in Section 6.3. The ranking is then refined with a supervised learning to rank model using RankLib⁴. The final step is NIL handling which determines if the mention is in the knowledge base or whether it is unknown.

⁴<http://cs.umass.edu/~vdang/ranklib.html>

6.6.1 Knowledge Base Representation

Our system addresses text-driven knowledge bases in which each entity is associated with free text, with relationships between entities from hyperlinks or other sources. Wikipedia is one representation of such a knowledge base, but our system would likely perform well on other knowledge bases.

In order to efficiently search over knowledge bases with millions or billions of entities we use an information retrieval system. For these experiments, we index the full text of the Wikipedia article, title, redirects, Freebase name variations, internal anchor text, and web anchor text.

6.6.2 Document Analysis

The first step in linking is to identify the entity query span q_m in the document and to find disambiguating contextual information for the query model introduced in Section 6.4. This includes name variations V , contextual sentences S , and other neighboring mentions M .

In the TAC KBP challenge, the entities of type person, organization, or location are the main focus of the linking effort and so the system detects entities using standard named entity recognition tools, including UMass' *factorie*⁵ and Stanford CoreNLP⁶. These provide the mentions spans to derive the name span q , name variations V , and neighboring entities M . Beyond the standard entity classes, our approach is general enough to also link other entity types if a suitable detector is incorporated.

Given a target entity mention, m , the system needs to identify name variations, V , in the document, such as “Steve” to “Steve Jobs” or “IOC” to “International Olympic Committee”. The goal is to identify alternative names that are less ambiguous than the query mention. We use the within-document coreference tool from UMass' *factorie*, together with capitalized word sequences that contain the query string (ignoring capitalization and punctuation for

⁵<http://factorie.cs.umass.edu/>

⁶<http://nlp.stanford.edu/software/corenlp.shtml>

Feature Set	Type	Description
Character Similarity	q, v	Lower-cased normalized string similarity: Exact match, prefix match, Dice, Jaccard, Levenshtein, Jaro-Winkler
Token Similarity	q,v	Lower-cased normalized token similarity: Exact match, Dice, Jaccard
Acronym match	q	Tests if query is an acronym, if first letters match, and if KB entry name is a possible acronym expansion
Field matches	q, v	Field counts and query likelihood probabilities for title, anchor text, redirects, alternative names fields
Link Probability	q, v	p (anchor — KB entry) - the fraction of internal and external total anchor strings targeting the entity
Inlink count	document prior	Log of the number of internal and external links to the target KB entry
Text Similarity	document	Normalized text similarity of document and KB entity: Cosine with TF-IDF, KL, JS, Jaccard token overlap
Neighborhood text similarity	document	Normalized neighborhood similarity: KL Divergence, Number of matches, match probability
Neighborhood link similarity	document	Neighborhood similarity with in/out links: KL divergence, Jensen-Shannon Divergence, Dice overlap, Jaccard
Rank features	retrieval	Raw retrieval log likelihood, Normalized posterior probability, 1/retrieval_rank
Context Rank Features	retrieval	retrieval scores for each contextual components: q, v, s, m_nrm, m_local

Table 6.1: Features of the query mention and candidate Wikipedia entity.

the matching) to extract name variations V . From the set of coreferent mentions, we extract the sentences S they occur within. After removing stopwords, casing and punctuation they represent non-NER context such as verbs, adjectives, and multi-word phrases.

6.6.3 KB Entity Ranking

The query model with salience weights from local document analysis and the neighborhood relevance model $\rho^{\text{nrm}}(m)$ is executed against the search index of KB entries as shown in Equation 6.2. Our system supports any feature function expressible in Galago query notation. Beyond this initial ranking, we can further refine the ranking using more complex features in learning to rank models.

The ranking is refined using supervised machine learning in a learning to rank (LTR) model. The refinement employs more extensive feature comparisons which would be expensive to compute over the entire collection. For these experiments we use the LambdaMART ranking model, a type of gradient boosted decision tree that is state-of-the-art and captures non-linear dependencies in the data. The model includes dozens of features. A description of the features used in the model is found in Table 6.1.

6.6.4 NIL Handling

After the entities are ranked, the last step is to determine if the top-ranked entity for a mention is correct and should be linked to the KB entry or instead refers to an entity not in the knowledge base, in which case NIL should be returned. For these experiments, we use a simple NIL handling strategy. We return NIL if the supervised score of the top ranked entity is below a threshold τ . The NIL threshold τ is tuned on the training data. For the special case of the TAC KBP challenge, the reference knowledge base is a subset of Wikipedia. We exploit this fact by returning NIL whenever the top ranked Wikipedia entity is not contained in the reference knowledge base.

6.7 Experimental Evaluation

6.7.1 Target Model Evaluation

We first evaluate the contributions of the different types of local document context for the entity query. The context includes the entity query string q , name variations v , the sentences s surrounding the query or name variations, as well as neighboring entity spans m . The combinations of these context components is indicated by Q, V, S, or M in the method prefix.

We evaluate three context weighting methods. The first is uniform weighting (QVM). Second is the local document model by Gottipati and Jiang (GOTTIPATI and JIANG 2011) (indicated by local). Third is our neighborhood relevance model (indicated by the suffix nrm). We compare both for estimating the salience $\rho(m)$ of neighboring entity mentions m . Baselines are the methods using only the query string (Q), the combination of query and name variations (QV), as well as the local context weighting (QVM_local). Our suggested methods are QVSM_nrm and QVM_nrm. These models are the full query model with neighborhood relevance weighting with and without sentences.

For each of the compared methods, we train separate λ parameters on the training data using a coordinate ascent learning algorithm. Estimated λ parameters differ across methods.

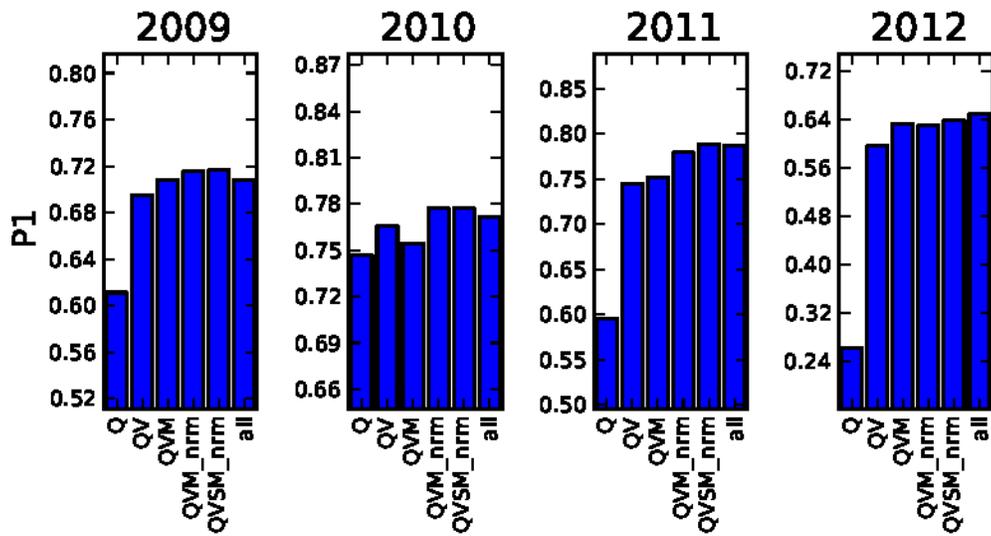
For the QVSM_nrm model the estimated parameters are: $\lambda^Q = 0.321$, $\lambda^V = 0.293$, $\lambda^S = 0.155$, and $\lambda^M = 0.230$.

Figure 6.4 visualizes an ablation study for the context components using precision at rank one for evaluation. Figure 6.4a shows the cumulative improvements as context is added and weighted with the neighborhood relevance (QVM_nrm). QVM with uniform neighborhood weighting performs similarly to QVM_local weighting (not shown). We observe that adding sentence context does not significantly improve performance.

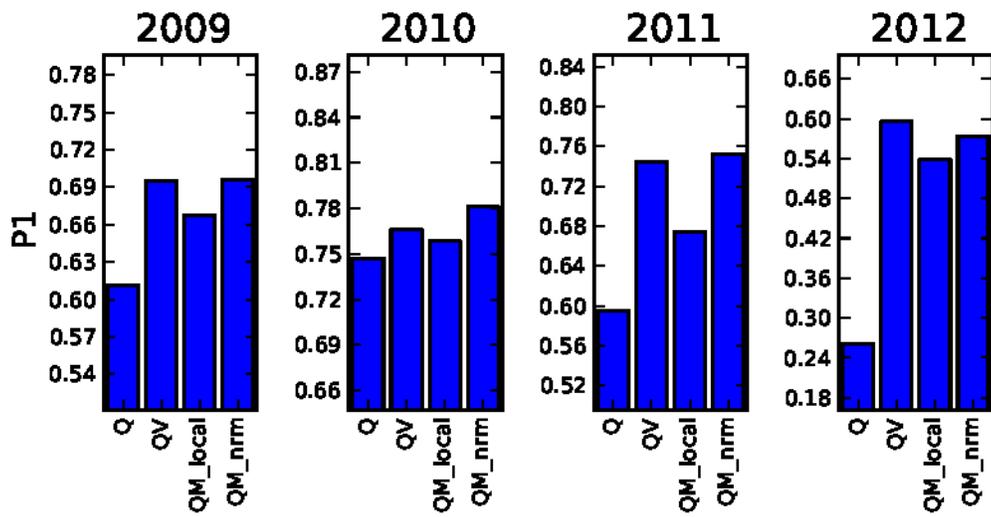
Figure 6.4b details the individual contributions of contextual components (omitting sentences). It is interesting that the QV method (entity name plus name variations) and QM_nrm (entity name plus weighted entity spans) are comparable in effectiveness. This is useful when no high quality name variations are extractable from the text, as is the case in informal text from social media. The cumulative figure above shows that when combined these features yield further improvement. Across all years the neighborhood relevance model achieves better effectiveness than the local model.

6.7.2 Ranking Evaluation

In the previous section we examined the effectiveness of the different contextual components only on the top-ranked result. Table 6.2 presents the contextual ranking models evaluated using mean reciprocal rank (MRR). Similar to the previous results, it shows that the most effective models include the neighborhood relevance weighting scheme (nrm). QVM_nrm and QVSM_nrm are significantly better than the QV baseline. The only exception is in 2010, when the queries are easier. In this case only the QM_nrm method is significantly better. Additionally, QVM_nrm is significantly better than the local weighting (Gottipati) for 2009-2011. However, there is no significant difference in 2012. We hypothesize that the reason the neighborhood model does not improve over the local model in 2012 is because the queries are significantly more ambiguous and the quality of the retrieved feedback documents is lower.



(a) Cumulative.



(b) Individual Contributions.

Figure 6.4: Ablation study for the suggested method in terms of Precision @ 1.

Method	2009	2010	2011	2012
Q	0.702	0.824	0.698	0.385
QV	0.772	0.838	0.821	0.686
QM_nrm	0.773	0.849*	0.825*	0.666
QM	0.746	0.829	0.758	0.636
QVM_nrm	0.795*	0.845	0.849*	0.715*
QVM_local	0.784*	0.829	0.831	0.730*
QVS	0.771	0.834	0.822	0.697*
QVSM_nrm	0.792*	0.845	0.850*	0.726*
QVSM_local	0.780*	0.836	0.837*	0.719*
all context	0.786*	0.841	0.848*	0.735*

Table 6.2: Ranking results on TAC by year with varying context methods with mean reciprocal rank (MRR). The best results for each year are highlighted in bold. Results that are statistically significant with $\alpha = 5\%$ over the QV baseline are indicated with *.

Method	2009	2010	2011	2012
QVM_nrm	0.795	0.845	0.849	0.715
QVM_nrm LTR	0.913	0.936	0.918	0.805

Table 6.3: Learning to rank refinement results with mean reciprocal rank (MRR). All LTR results are statistically significant with $\alpha = 5\%$ over the unsupervised QVM_nrm

We refine the retrieval ranking using a supervised learning to rank model. The the features in the ranking model are described in Table 6.1. The top 100 results from the best ranking, QVM_nrm are reranked. The results of this are shown in Table 6.3. The results show significant improvement over the initial retrieval ranking leveraging more features that perform more extensive contextual comparison. The results for 2012 are still well below the other years, indicating the difficulty of these queries even leveraging the more complex contextual features. This indicates that a better feature representation is needed to address some of these difficult to resolve mentions.

6.7.2.1 Recall

The previous results use mean reciprocal rank to measure the retrieval effectiveness. We now examine the rank distribution in more detail, examining the recall at a given rank cutoff. The entity recall is critical because it is an upper bound on the effectiveness of downstream

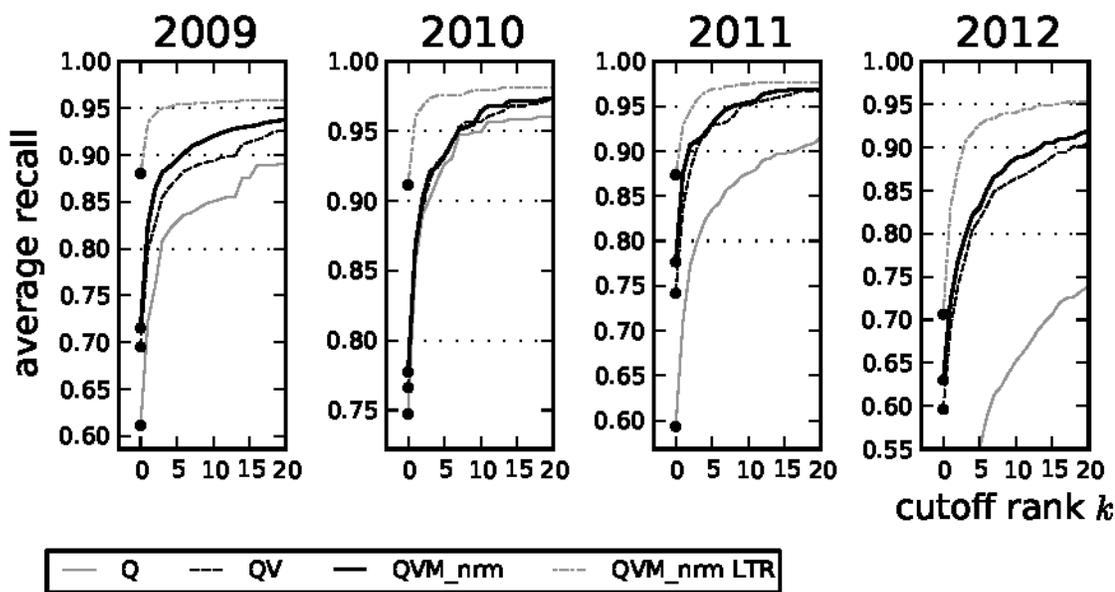


Figure 6.5: Average recall at rank cutoff k .

	2009			2010			2011			2012		
	in-KB	NIL	all									
QVM_nrm	0.810	0.703	0.764	0.768	0.764	0.766	0.766	0.767	0.766	0.584	0.623	0.605
QVM_nrm LTR	0.861	0.798	0.825	0.892	0.762	0.822	0.858	0.756	0.805	0.705	0.628	0.668
QVRM_nrm LTR NIL	0.847	0.848	0.847	0.883	0.843	0.862	0.833	0.857	0.845	0.676	0.758	0.714
Best Performer	0.765	-	0.822	0.823	-	0.864	0.801	-	0.870	0.687	-	0.721

Table 6.4: TAC Entity Linking performance in macro-average accuracy.

systems. To achieve a minimum 90% recall threshold across all years requires hundreds of candidates for the query (Q) model, 20 for QV, 16 for QVM_nrm, and only 3 for QVM_nrm LTR. The learning to rank model achieves at least 95% recall across all years within 10 results.

6.7.3 TAC KBP results

In this section, we evaluate the ranking as part of the entire linking pipeline described in Section 6.6. We report the macro-averaged accuracy because we do not focus on clustering NIL entity mentions. The results are in Table 6.4. The unsupervised retrieval QVM_nrm

performs well, above the median in 2012 and competitive in previous years. The supervised ranking models improve effectiveness significantly. The in-KB ranking results outperform the best performing systems in 2009 through 2011 and are comparable in 2012. The main focus of this work is ranking, and this shows the effectiveness of our approach.

We now examine the overall (all) results, including the NIL handling. The results show that the QVM_nrm with LTR reranking and NIL handling outperforms the top system in 2009 and is competitive with the best performing systems in subsequent years. Applying the score threshold improves the overall accuracy despite decreasing in-KB effectiveness. This is because some correctly linked entities are marked as NIL, but are outweighed by the greater reduction in false positive entity links. The NIL handling strategy based on thresholding the ranking score is effective, but could be improved further. Other linking systems use a supervised NIL classifier for this step, allowing them to perform well despite less effective in-KB ranking.

6.8 Enrichment approaches for TAC KBP 2013

In this section we describe further experiments conducted on the TAC KBP 2013 entity linking task ⁷. This section is based upon the official TAC submission for 2013. We evaluated two new enrichment approaches for 2013. The first is enrichment using Urban Dictionary ⁸. For mentions in forum documents, we search Urban Dictionary for the mention string and perform enrichment with returned entity tags. The second method is a multi-pass linking model. Instead of linking only the TAC query mention, all entity mentions in the document are linked to the knowledge base using features that encourage a coherent assignment of the linked entities.

⁷This section is partially based upon work published at Text Analysis Conference (TAC) Knowledge Base Population (KBP) entity linking track (TAC KBP '13) (DALTON and DIETZ 2013c).

⁸<http://www.urbandictionary.com/>

6.8.1 Urban Dictionary Enrichment

In TAC 2013 a significant fraction of the mentions in the evaluation set are from forum data. Unlike previous newswire and web data, the mentions in this data appear to contain a high fraction of creative slang and pop culture terms. Examples of such slang query mentions include: [McSame], [MCCane], [Biebs], [Obamessiah], [Nobama], [Turd Blossom], and [uz-becky-becky-becky-stan-stan]. The existing sources of aliases from anchor text and structured metadata are unlikely to contain these informal references.

To address the vocabulary mismatch, we use Urban Dictionary as an external source for enrichment. Urban Dictionary is a crowd-sourced online web dictionary with more than seven million definitions, focused on slang and pop culture phrases not found in standard dictionaries. For example, [McSame] has the definition: “John McCain. He considers himself a straight talking maverick, when in reality he is merely running on the promise of four more years of George W. Bush.” We leverage the entries as a source for mention enrichment to include in the retrieval context.

To perform enrichment using urban dictionary, we perform source retrieval against the urban dictionary web service to retrieve definitions. We retrieve the full-text of the definition, as well as the tags. For simplicity, we focused on the article tags, which often include the name of the entity being described. These tags were added to the neighboring entities m and consequently as part of the retrieval query. The main goal of this step is to improve the recall of our retrieved entities from the knowledge base.

6.8.2 Entity KB Coherence

We also experiment using fully disambiguated document representations. In this model, entity extraction and disambiguation is performed on all entities in the document in a first pass. Then, a second step that leverages the features from disambiguated mentions re-ranks the possible links with a model that includes entity-to-entity compatibility features. This

Feature Set	Description
Category IDs	Intersection, Misses, Dice, Jaccard, Cosine
Category Words	Jaccard, Jensen-Shannon Divergence, Cosine with TF-IDF, Unweighted cosine
Article Text	Jaccard, Jensen-Shannon Divergence, Cosine similarity
Text Mentions	Contains entity name, Both articles contain name
Inlinks	Pointwise mutual information, ProxPMI (wikifier), Intersection, Jaccard, Dice, Google Norm. Distance
Outlinks	Pointwise mutual information, ProxPMI (wikifier), Intersection, Jaccard, Dice, Google Norm. Distance
Inlinks + Outlinks	Pointwise mutual information, ProxPMI (wikifier), Intersection, Jaccard, Dice, Google Norm. Distance
Shared Links	Linked, mutal link

Table 6.5: Features of the entity-to-entity similarity.

is a joint assignment model similar to the techniques employed by other leading systems (CUCERZAN 2011; LEHMANN *et al.* 2010; RATINOV *et al.* 2011).

A recent trend in entity linking has been joint or ‘collective’ assignment of mentions in a document. The HLTCOE introduced the Context Aware Linker of Entities (CALE) using local context entities (STOYANOV *et al.* 2012). Language Computer Corporation (LCC) uses features from a subset of the closest unambiguous mentions (LEHMANN *et al.* 2010). The Microsoft system for TAC builds a context vector from the union of candidates for all entities (CUCERZAN 2011). UIUC’s GLOW system uses ‘global’ similarity features from a first pass linking model (RATINOV *et al.* 2011).

We implemented an extension to our supervised ranking model that incorporates features similar to Wikifier. We first perform a first pass ranking, taking all mentions that would be predicted as non-NIL as context links. For documents with large numbers of entities, we limited the context to the 50 links with the highest compatibility score. We use the features described by GLOW as well as those from MSR. A full list of the features are given in the Table 6.5.

Although this model proved promising on training data, we found that the supervised model did not generalize well to the 2013 data distribution. We hypothesize that it did not perform as well as expected because of limited training data.

Approach	Run ID	Acc.	B ³⁺ Prec.	B ³⁺ Recall	B ³⁺ F1
QVM	UMass_CIIIR1	0.577	0.573	0.317	0.408
QVM LTR	UMass_CIIIR2	0.729	0.716	0.462	0.561
QV LTR	UMass_CIIIR3	0.802	0.781	0.571	0.660
QVM UrbDict LTR	UMass_CIIIR4	0.806	0.785	0.584	0.670
QVM E2E LTR	UMass_CIIIR5	0.746	0.730	0.503	0.595
2013 Median		0.746	0.718	0.496	0.574
2013 Best		0.833	0.826	0.689	0.746

Table 6.6: Overall effectiveness in 2013.

Approach short description	Run ID	News	Web	Forum
QVM	UMass_CIIIR1	0.493	0.528	0.202
QVM LTR	UMass_CIIIR2	0.637	0.609	0.414
QV LTR	UMass_CIIIR3	0.743	0.615	0.547
QVM UrbDict LTR	UMass_CIIIR4	0.745	0.620	0.572
QVM E2E LTR	UMass_CIIIR5	0.667	0.638	0.457
2013 Median		0.645	0.525	0.488
2013 Best		0.829	0.678	0.662

Table 6.7: B³⁺ F1 by document type.

6.8.3 Results

The overall results of our runs are shown in Table 6.6. The results by document type are in Table 6.7 and by entity class in Table 6.8. Unlike the results in 2012, the unsupervised retrieval model, UMass_CIIIR1, performed significantly below the median, especially on the forum data with a B³⁺ F1 value of only 0.202. It also struggled with GPE entities with a B³⁺ F1 of only 0.091. However, it performs above the median on ORGs. It is clear that more effective context models are needed for both GPEs and forum data.

	QVM	QVM LTR	QV LTR	QVM UrbDict LTR	QVM E2E LTR	Median	Best
PER	0.576	0.671	0.694	0.709	0.722	0.627	0.778
ORG	0.590	0.638	0.626	0.639	0.662	0.542	0.737
GPE	0.091	0.399	0.657	0.660	0.424	0.552	0.746

Table 6.8: B³⁺ F1 by entity class.

6.9 Summary

In this chapter we introduced a retrieval-based method for the task of linking detected entity mentions to an external knowledge base. We use an entity-based enrichment method, the *Neighborhood Relevance Model*, which focuses on identifying salient associations between an entity mention and other entities in the local document neighborhood. The neighborhood relevance model uses the pattern of entity mentions in similar sources to identify salient entity context.

Our experiments on the TAC KBP entity linking data show that this enrichment model outperforms other context weighting models. The results show up to a 16.4% improvement in mean reciprocal rank over local models for entity linking (Contribution 2). When combined with a learning to rank model that incorporates more text similarity features, the results beat the current best performing systems on in-KB ranking accuracy. Combined with a simple NIL handling strategy the overall effectiveness on all mentions is comparable to, and sometimes better than, other state-of-the-art entity linking systems.

CHAPTER 7

ENTITY-BASED FEATURE ENRICHMENT FOR RETRIEVAL

In the previous two chapters, we built up increasingly rich entity representations of documents using information extraction, detecting and disambiguating named entities. In this chapter we focus on the task of ad hoc document retrieval. We apply the enrichment framework to queries and documents, expanding them with structured and unstructured features from entity mentions. We demonstrate that enriching the query representation using entity features results in significant improvements in retrieval effectiveness.¹

This chapter focuses on two research areas using entity annotations for ad hoc retrieval. The first is the representation of both queries and documents with linked entities. What features, if any, improve retrieval effectiveness? The second is how to infer latent entities (and more importantly, features of entities and terms) for an information need.

Using similar automatic extraction methods to those described in chapters 5 and 6, leading web search companies are extracting and linking entities in text web documents. To recap a description provided earlier, Google recently released the FACC1 dataset (GABRILOVICH *et al.* 2013) for the TREC ClueWeb09 and ClueWeb12 web collections. The dataset contains automatically extracted entity mentions from web documents linkable to the Freebase knowledge base (BOLLACKER *et al.* 2008). Freebase is a publicly available general purpose knowledge base with over 42 million entities and over 2.3 billion facts.² The FACC1 dataset is the first publicly available web-scale collection of entity linked documents. In addition

¹This chapter is partially based upon work published at the 37th Annual ACM Special Interest Group for Information Retrieval (SIGIR '14) (DALTON *et al.* 2014).

²As of January 27, 2014 according to Freebase.com

to annotated documents, the FACC1 data also contains explicit manual annotations for the TREC web track queries. We present the first published experiments using this data for retrieval, to our knowledge.

The FACC1 ClueWeb annotations include entity annotations for queries. However, these annotations are limited to entities that are explicitly mentioned, where we hypothesize that many more latent entities are relevant to the users' information need. Similar to word-based models, explicit entity representations of queries have fundamental problems of query-document mismatch. In addition, many existing text collections do not have explicit entity annotations for queries. For both of these cases, we can leverage expansion models to expand the entity representation.

Entities provide a wealth of rich features that could be used for representation. The features include text as well as structured data. Some of the important attributes that we highlight for these experiments include: fine-grained type information (athlete, museum, restaurant), category classifications, and associations to other entities. Although we do not explore them in detail in this work we also observe that the knowledge base contains rich relational data with attributes and relations to other entities. For different types of entities these attributes include: gender, nationality, profession, geographical information (latitude, longitude), and temporal attributes (such as birth and death).

We hypothesize that the language used to describe entities in the document collections differs from that found in Wikipedia or in the knowledge base description. We propose query-specific entity context models extracted from snippets in the feedback documents surrounding the entity's annotations to model entities from the collection. This is related to previous work on local expansion models of concepts (XU and CROFT 1996), but we perform it on disambiguated entities. We use these entity context models to rank entities with respect to the query as well as extract expansion features.

This chapter includes all of the contributions listed in Chapter 1. It contains three contributions unique to this chapter:

1. **We present the first known experimental results using entity linked documents and queries for ad hoc document retrieval.** We experiment using linked entities for newswire and web test collections. We use documents annotated with linked entities provided by the KB Bridge entity linking system and the openly available FACC1 entity annotations by Google for web data. For a subset of these collections we also experiment with entity linked queries. We compare models incorporating entity features with state-of-the-art word-based models. Compared with competitive query expansion baselines, the sequential dependence model with relevance modeling expansion on Wikipedia and the collection, there is an improvement of 16.4% and 11.5% in MAP on Robust04 and a 14.1% and 32.8% improvement in NDCG@20 for ClueWeb12. For ClueWeb09, where results do not significantly improve, we perform an error analysis and identify several important underlying causes for this behavior.
2. **We define a new query-specific entity context model that models the feature context of an entity using retrieved documents.** Existing entity context models are built globally across a collection. Previous local models (XU and CROFT 1996) use word and phrase features from noun phrases that are not disambiguated for expansion. We introduce a new query-specific entity model that includes both words, phrases, and features from disambiguated entity mentions. For the task of retrieval, we show that these models provide an effective mechanism for identifying the relevance of entities and as a source of expansion features.
3. **We extend dependency models to include entity-based features that model dependencies between text and different types of structured entity features.** Existing query expansion techniques (METZLER and CROFT 2007) model dependencies derived from words, including phrase and proximity concepts. We propose a feature expansion model that models dependencies between text and structured entity features including: entities, fine-grained types, categories, and entity associations. For

example a dependency between a type of entity and a word: (/boats/ship sinking) or (/government/politician scandal).

7.1 Entity-Based Document Retrieval

In this section we describe the retrieval model we use to retrieve documents using linked entities. For these experiments we use proven retrieve models and vary the representation of the queries and documents used. Specifically, we use the query likelihood and sequential dependence retrieval models described earlier in Chapter 2. We now describe the document and query representation incorporating entity links.

The user query Q is given as a sequence of observed keywords $w_1w_2, \dots w_{|Q|}$. The queries may also be labeled with entity mentions M_Q . Each mention, m contains its text, as well as a multinomial distribution over links to entities in the knowledge base $e \in E$ (cf. Figure 7.4a)

For the document representation, we are given a text document containing a sequence of words, $w_1w_2, \dots w_{|d|}$. The text documents are also annotated with entity mentions, M_d linked to knowledge base entities. Similar to queries, each mention, $m \in M_d$, has a multinomial distribution over possible entities, $e \in E$.

For both queries and documents the entity links establish a bidirectional link from text mentions to entities in the KB, and through these links to structured attributes such as Freebase types and Wikipedia categories. The knowledge base also contains associations and relations to other entities, E' (which we refer to as neighbors). We refer to each of these different attribute types as *vocabularies*, V , over which counts and matches can be computed. The vocabularies we use in these experiments are shown in Figure 7.4.

Just as we index words with their position in the document, we similarly index entities and their attributes with their corresponding word positions. For scoring, the inverted index structure provides access to counts and positions of the different vocabulary elements for the



Topic

Barack Obama ^{en}

[mjr:/m/02mjmr](#) · [notable type: government_bus_president](#) · [notable for: government_bus_president](#) · [on the web: !\[\]\(f9cd3d0fa7c31b255f85f1032880eb06_img.jpg\) \[wikidata.org\]\(#\)](#) · [Created by \[User:Wb\]\(#\) on 10/22/2006](#)

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 the 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. In 2004, Obama received national attention during his campaign to represent Illinois in the United States Senate, with his victory in the March Democratic Party primary, his keynote address at the Democratic National Convention in July, and his election to the Senate in November. He began his presidential campaign in 2007 and, after a close primary campaign against Hillary Clinton, won in 2008, he won sufficient delegates in the Democratic Party primaries to receive the presidential nomination. He then defeated Republican nominee John McCain in the general election, and was inaugurated as president on January 20, 2009. Nine months after his election, Obama was named the 2009 Nobel Peace Prize laureate. [..]

Properties

ISBN

Keys

Links



View and edit specific domains, types, or properties...

Filter options: Show all domains and properties

Common actions

Topic actions

Also known as actions

Barack Hussein Obama II

Barack Hussein Obama, Jr.

Barack Hussein Obama

Obama

President Obama

Barack H. Obama II

Barack H. Obama III

Barack Obama II

President Barack Hussein Obama II

Sen. Barack Obama

Barak Obama

[76 values total >](#)

Description actions

Barack Hussein Obama II (Barak Hussein Obama; born August 4, 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama served as a U.S. Senator representing the state of Illinois from January 2005 to November 2008, when he resigned following his victory in the 2008 presidential election. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the 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 in Chicago and taught constitutional law at the 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.

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 in Chicago and taught constitutional law at the 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. In 2004, Obama received national attention during his campaign to represent Illinois in the United States Senate with his victory in the March Democratic Party primary, his keynote address at the Democratic National Convention in July, and his election to the Senate in November. He began his presidential campaign in 2007 and, after a close primary campaign against Hillary Clinton in 2008, he won sufficient delegates in the Democratic Party primaries to receive the presidential nomination. He then defeated Republican nominee John McCain in the general election, and was inaugurated as president on January 20, 2009. Nine months after his election, Obama was named the 2009 Nobel Peace Prize laureate. [..]

[42 values total >](#)

Image actions





Official website actions

<http://www.barackobama.com>

Types:

- Metaweb System Types
- Object
- Common
- Topic
- Identity
- Education
- Honorary Degree Recipient
- Film
- Film subject
- Person or entity appearing in film
- Governmental
- US President
- Politician
- Political Appointee
- U.S. Congressperson
- Honored entity
- Robert's types
- Presidential Candidate
- My favorite things
- Daylife
- collin's types
- Twitter Topic
- Music
- Musical Artist
- Musician
- Composer
- Business
- Employer
- Architecture
- Building Occupant
- Books
- Literature Subject
- Poem character
- TV
- TV Actor
- TV Personality

Figure 7.1: Example Freebase entity for Barack Obama, /m/02mjmr

Mention	Start byte	End byte	Posterior	Freebase ID	Entity Name
GOP	19722	19725	0.992	/m/07wbk	Republican_Party_(United_States)
Albuquerque	24759	24770	0.989	/m/0djd3	Albuquerque,_New_Mexico
Barack Obama	24876	24888	0.996	/m/02mjmr	Barack_Obama
Senator Obama	24945	24958	0.996	/m/02mjmr	Barack_Obama
Obama	25093	25098	0.996	/m/02mjmr	Barack_Obama

Figure 7.2: FACC1 entity document annotations for clueweb09-en0004-08-20390

mention	start byte	end byte	freebase ID	posterior
obama	0	10	/m/02mjmr	0.998
family tree	12	22	/m/016p0k	0.806

Figure 7.3: FACC1 entity annotations for TREC Web Track query 1: [obama family tree]

documents. This representation allows us to model dependencies across vocabulary types, for example a category {Politician} near the word “family”.

7.1.1 Cross-vocabulary dependencies

Previous models (METZLER and CROFT 2005; METZLER and CROFT 2007; BENDER-SKY and CROFT 2012) generate dependencies using only word-derived concepts. In this work, we introduce dependencies from entity features. We derive the features within a given vocabulary (an entity near another entity) and across-vocabularies (a category near a word). This is possible because the vocabulary information is positional with respect to query word vector, W . An example query incorporating these dependencies is shown in Figure 7.5. For example, entity type and word dependencies, such as t_1 and w_2 in the example above are represented by the feature $f_{\text{ExpIType}}(Q, “T, W”)$ for a particular type-word dependency “ t_1, w_2 ” is computed by integrating over multiple occurrences $p(“M, w_2”|Q)$ and integrating over conditional entity and type distributions.

$$f_{\text{ExpIType}}(Q, “T, W”) = \left(\sum_{m \in M} \left(\sum_{e \in E} p(t|e)p(e|m) \right) p(“m, w”|Q) \right) p(W = w|Q)$$

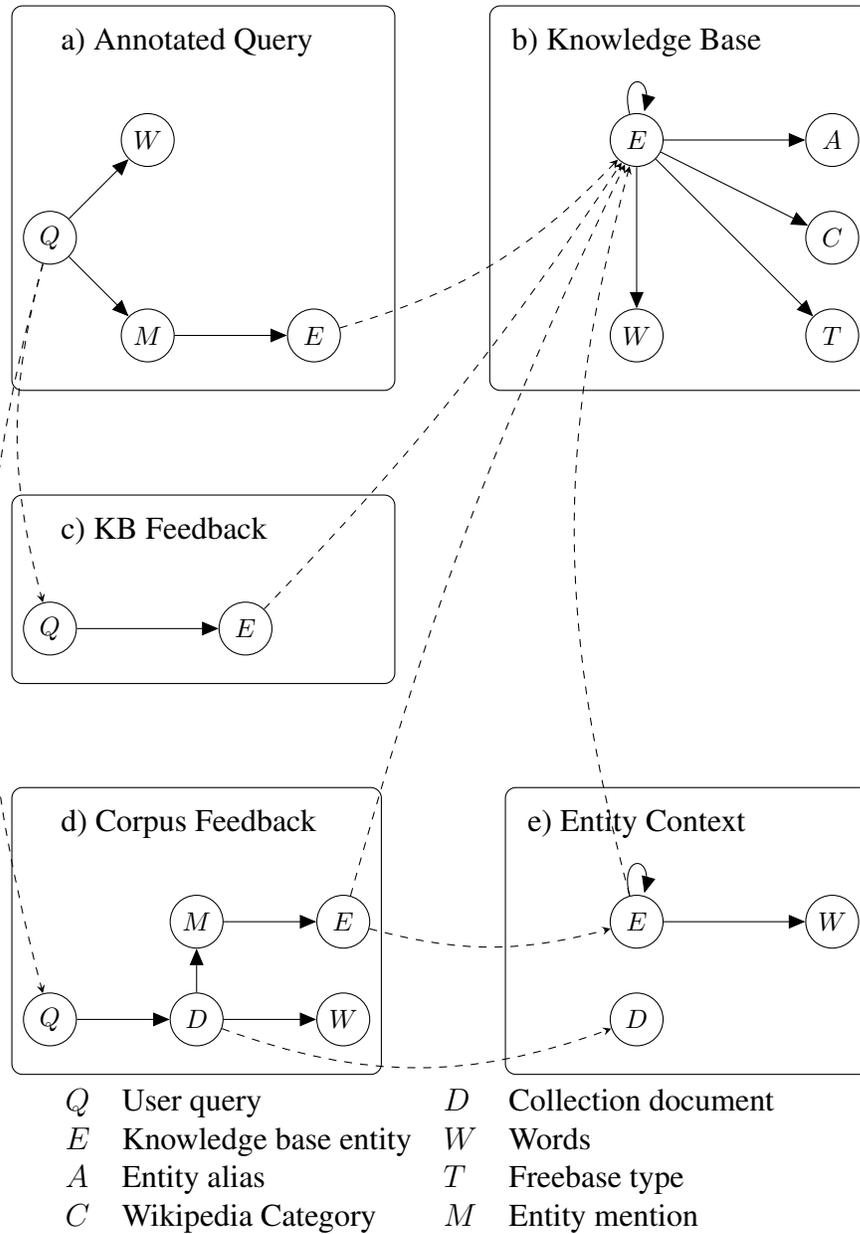


Figure 7.4: Overview over feature sources.

```

#combine(
  #sdm( obama family tree )
  #sdm( [Barack_Obama] family tree)
  #sdm( obama [Family_Tree])
  #sdm( [Barack_Obama] [Family_Tree] )
  #sdm( {US President} family tree )
  #sdm( {Politician} family tree )
)

```

Figure 7.5: Example expansion of query C09-1 with entities [] and Freebase types {}.

The equation uses entity disambiguation confidences $p(e|m)$ and entity-specific type distributions $p(t|e)$.

In addition to structured attributions, terms on the entity’s knowledge base entry also provide a resource for related words W . These are derived through a hierarchical multinomial model by integrating over mentions and entities

$$f_{\text{ExpWiki}}(Q, W) = \sum_{m \in M} \left(\sum_{e \in E} p(W|e)p(e|m) \right) p(m|Q)$$

In the equation $p(m|Q)$ is a uniform distribution over annotated mentions and $p(e|m)$ is the entity disambiguation confidence and $p(W|e)$ refers to the language model of the entity’s knowledge base entry.

7.2 Entity Context Model

In the previous section, we described a documents incorporating entity features available in the knowledge base. These features include the language of the entity, $p(W|e)$, and associations to other entities. While this information is useful, it is not the complete picture of an entity. In particular, the language and associations of an encyclopedia may be very different the language and associations found in documents. To address this, we propose deriving a new representation of an entity from the documents in the text collection.

We introduce a new local model of an entity created from mentions, M_D in retrieved documents (Contribution 3). We construct an entity pseudo-document for each unique disambiguated entity, e . For each entity, we extract local context information from the words and entities that co-occur within a particular scope near the entity mention. We aggregate all the local context snippets for each entity, including weighting the snippets by the mention source retrieval probability $p(d|Q)$ and the probability of the entity for the mention $p(e|m)$. These two factors incorporate the similarity to the query and the certainty of disambiguation in modeling the feature representation of the entity.

To extract the features from a document, we inspect the local context surrounding entity mention, $m \in M_d$ to extract sequences of words and other entity mentions. In our experiments, we create three versions of each entity’s model, varying the size of the context snippets: 8 words on either side of a mention, 50 words on either side, or one sentence, where sentence boundaries are determined by a sentence-splitter.

The local context model that we propose has several important differences from previous entity modeling approaches. First, it uses disambiguated entity links rather than simple string matches for co-reference resolution. If there are multiple ambiguous mentions of the same name, the contexts are separated based on their linked entity. Also, we do this for all types of concepts that exist in the knowledge base rather than just traditional *named* entities (person, organization, location).

Second, our context models are query focused. We construct an ECM from documents retrieved in response to the query. This change is important for large corpora because for entities with multiple diverse topics a representation across the entire collection will blend the topics together and lose their distinguishing characteristics. For example, the clueweb09 query [obama family tree] focuses on aspects of Obama’s family life and relationships to relatives, which is a relatively obscure topic when compared with more popular aspects such as “obamacare”.

The proposed entity context model captures not just words and phrases co-occurrence counts, but weighted the snippets preserve word and entity order surrounding the mention. The feature representation of an entity pseudo-document includes structured entity features from co-occurring entities: their mentions and features of them, including types and categories.

7.2.1 Related entity models

Building entity context models from their surrounding representation has been studied in the past. In 1994, Conrad and Utt (CONRAD and UTT 1994) used all paragraphs in the corpus surrounding *named* entity mentions to represent the entity, allowing free text queries to find names associated with a query. Ten years later, Raghavan et al. (RAGHAVAN *et al.* 2004) extended that idea to use language modeling as a representation and showed that these models could successfully be used to cluster, classify, or answer questions about entities. In these cases, the entity’s context was a paragraph or a fixed number of words surrounding all mentions of the entity in the corpus. More recently, the work of Schlaefter et al. (SCHLAEFER *et al.* 2011) expanded the representation of a Wikipedia entity using extracted “text nuggets” from the web for use in the Watson question answering system. Nuggets that were scored as relevant to the entity were used as its context, even if the nugget did not contain an actual mention.

Jing and Croft (1994) propose an approach, called PhraseFinder which builds an association thesaurus for a collection. They use noun phrases as the unit of text feature and associate words within a paragraph. The models are built across the entire collection. Later, in the context of expert search, Petkova and Croft (2007) associate named entities with text segments to build language models of person entities.

7.3 Entity Query Feature Enrichment

The goal is to derive an enriched query representation across the different kinds of vocabularies such as words W , entities E , types T and categories C to retrieve annotated documents with the goal of maximizing document retrieval effectiveness.

Figure 7.5 shows an example of expansions for the ClueWeb09B query 1 “obama family tree” for the words, entities, and Freebase types. We explicitly include proximity dependencies across vocabularies such as type $t_1 = \text{“Politician”}$, followed by word $w_2 = \text{“family”}$, and word $w_3 = \text{“tree”}$. A sample of the expansion terms from our experiment on this query are given in Table 7.1.

Enrichment in different vocabularies can be derived through multiple options. For instance, expansion entities can be found using pseudo-relevance feedback on the entity annotations or alternatively by feedback from the entity knowledge base. Figure 7.4 gives an overview of all possibilities we explore in this work.

Query Triggering

We perform enrichment for all the queries in the evaluation dataset. We do not evaluate triggering mechanisms for this task.

Target Model Generation

Unlike previous chapters, the queries used for document retrieval are short web queries with limited local context. The model we use for enrichment is described in Section 7.1.

Mention Source Retrieval

We perform source retrieval over three different sources of mentions. The first is the document collection, the second is the knowledge base, and finally the entity context models derived from contexts in retrieved documents. This is shown in Figure 7.4.

Document Collection We can also directly retrieve documents, D from the collection. These documents contain entity annotations (cf. Figure 7.4d corpus feedback). We note that

as before, our model incorporates the inherent link uncertainty, $p(e|m)$. We can integrate over the possible links to arrive at final distributions. If only the most confident entity is available we define $p(e|m) = 1$ for the linked entity e and 0 otherwise. From the distribution over entities, we can follow the connection to the knowledge base (cf. Figure 7.4b) and derive distribution over name aliases, types, categories, and neighbor entities.

Knowledge Base The query can be issued against a search index containing of the knowledge base (cf. Figure 7.4c kb feedback). The result is a distribution over entities, encoded in the feature $f_{KB}(Q, e)$. From the ranking of entities, we can also infer their distribution over their name aliases, types, categories and other features.

Entity Context Model The last source of feedback features is the entity context models, entity pseudo-documents, (cf. Figure 7.4e entity context). The result of retrieval is a distribution over entities, encoded in the feature $f_{ECM}(Q, e)$. From the entity context models we can also infer distributions over name aliases, types, categories, and related entities.

Mention Feature Extraction

The features, f , for query expansion include features from each of the different vocabularies: words, entities, mentions, types, categories, and neighboring entities. For each mention, m the features are derived using various options to traverse available information sources, each representing a path in Figure 7.4.

For every feature f , we build the expansion model induced by this feature only. For example from $f_{RM}(Q, e)$ we build an expansion model over entities $p_{RM}(e)$ by normalizing across all entity candidates E .

Feature Aggregation

We use the two-step aggregation model described in our framework described earlier in Chapter 4. The first per-source aggregation follows Equation 4.1. The second aggregation is

a weighted combination of these models incorporating the similarity to the target mention, as in Equation 4.2.

7.4 Experimental Setup

This section details the tools and datasets used for our experiments. The retrieval experiments described in this section are implemented using Galago³, an open source search engine. The structured query language supports exact matching, phrases, and proximity matches needed for our retrieval models. A summary of the document collections used in these experiments was presented in Chapter 3. The corpora include both newswire (Robust04) and web pages (ClueWeb). During indexing and retrieval, both documents and query words are stemmed using the Krovetz stemmer (KROVETZ 1993). Stopword removal is performed on word features using the Lemur 418 word stop list. For the web collections, the stopword list is augmented with a small collection of web-specific stopwords, including “com”, “html”, and “www”. We use title queries which contain only a few keywords.

Across all collections, all retrieval and feedback model parameters are learned or tuned using 5-fold cross-validation. Instead of selecting a single number of feedback documents or entities, we include expansion feature models with different hyper-parameters and learn weighted combination of these along with other features. We include expansion features using 1, 10, and 20 feedback entities and documents. We optimize parameters θ with a coordinate-ascent learning algorithm provided in the open source learning-to-rank-framework RankLib.⁴ Parameters are optimized for retrieval effectiveness using the metric mean average precision (MAP).

Retrieval effectiveness is evaluated with a variety of standard measures, including mean average precision (MAP) at 1000. Because several of our collections involve web search,

³<http://www.lemurproject.org/galago.php>

⁴<http://people.cs.umass.edu/~vdang/ranklib.html>

Words	Entity ID	Wiki Categories	Freebase Type
family	Barack_Obama	cat:first_families_u.s.	/people/family
tree	Michelle_Obama	cat:political_families_u.s.	/book/book_subject
genealogy	Family_Tree	cat:bush_family	/location/country
surname	Family_Crest	cat:american_families_english	/film/film_subject
history	Barack_Hussein_Obama_Sr	cat:american_families_german	/base/presidentialpets/first_family
crest	Family_History	cat:business_families_u.s.	/base/webisphere/topic

Table 7.1: Example expansion terms for the query “Obama Family Tree”

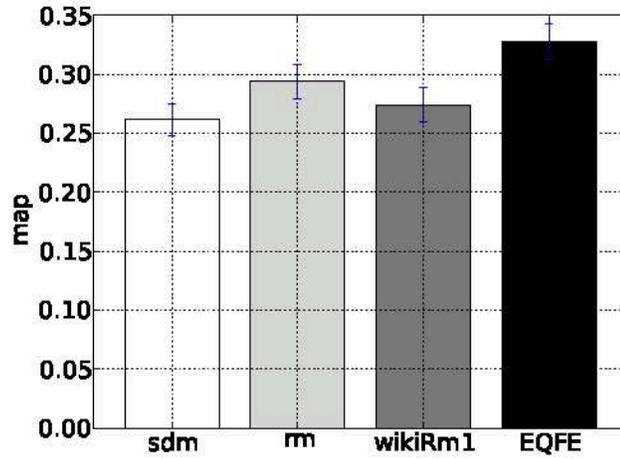
where precision at the early ranks is important, we also report normalized discounted cumulative gain (NCGD@20) and expected reciprocal rank (ERR@20).

7.5 Experimental Evaluation

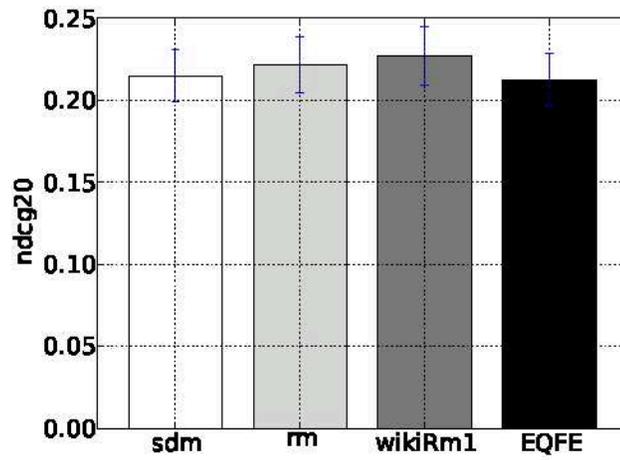
The effectiveness of our query feature expansion is compared with state-of-the-art word-based retrieval models and expansion alternatives. Our baseline model is the sequential dependence model. We use two baselines expansion models. The first is an external feedback model, which uses the Wikipedia knowledge base as a text collection and extracts terms from the top ranked article, which we call SDM-WikiRM-1. Models similar to SDM-WikiRM-1 were shown to be effective for these collections in previous work (BENDERSKY *et al.* 2012; XU *et al.* 2009). The second baseline uses collection ranking from the SDM model and builds a collection relevance model, which we call SDM-RM3. For ClueWeb12 we also report an official baseline using Indri’s query likelihood model (Indri-QL).

7.5.1 Overall Performance of EQFE

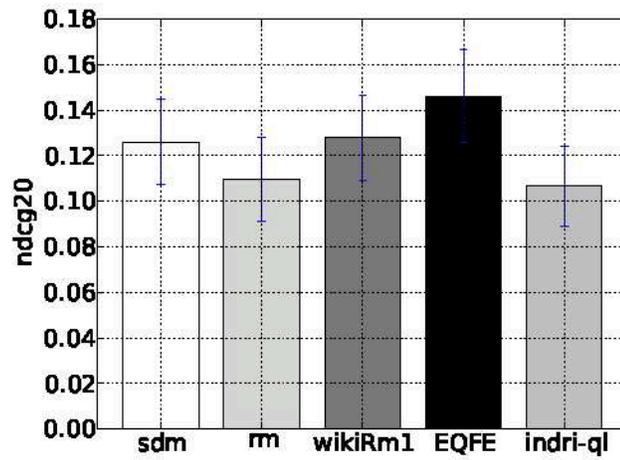
The overall retrieval effectiveness across different methods and collections is presented in Table 7.2 and Figure 7.6. Our EQFE model is best on MAP for Robust04 and best on NDCG@20, ERR@20 and MAP on ClueWeb12B. A paired-t-test with α -level 5% indicates that the improvement of EQFE over SDM (and the expansion models) is statistically significant. For ClueWeb09B, the EQFE numbers are slightly worse, but there is no significant difference detected among the competing methods.



(a) Robust04



(b) ClueWeb09B



(c) ClueWeb12B

Figure 7.6: Mean retrieval effectiveness with standard error bars.

Model	Robust04			ClueWeb09B			ClueWeb12B		
	MAP	P@20	NDCG@20	MAP	ERR@20	NDCG@20	MAP	ERR@20	NDCG@20
Indri-QL							3.64	07.79	10.66
SDM	26.15	37.52	42.37	11.43	13.63	21.40	4.18	09.15	12.61
SDM-WikiRM-1	27.41	37.71	42.81	11.39	15.29	22.56	4.00	09.31	12.80
SDM-RM3	29.38	38.82	43.44	11.43	13.63	21.40	3.53	07.61	11.00
EQFE	32.77	38.00	42.40	11.00	14.00	21.12	4.67	10.00	14.61

Table 7.2: Summary of results comparing EQFE with other methods across the three test collections.

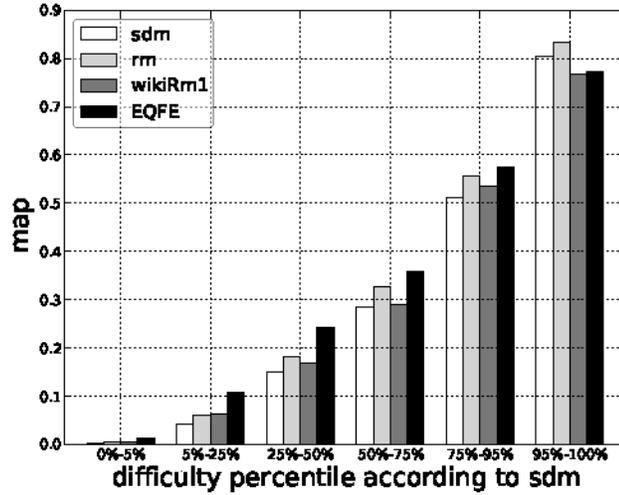
	Queries Helped	Queries Hurt
Robust04	173	47
ClueWeb09B	68	65
ClueWeb12B	26	8

Table 7.3: Queries EFQE helped versus hurt over SDM baseline.

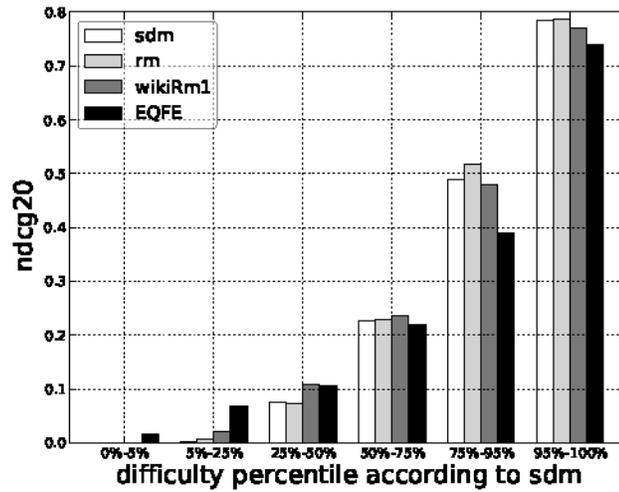
We perform a helps/hurts analysis for the methods in Table 7.3. The results show that more than three times the numbers of queries are helped than hurt for both Robust04 and ClueWeb12B. For ClueWeb09B, the results show that the method helps approximately the same number of queries that it helps.

In order to analyze whether the EQFE method particularly improves difficult or easy queries, we sub-divide each test set into percentiles according to the SDM baseline. In Figure 7.7 the 5% of the hardest queries are represented by the left-most cluster of columns, the 5% of the easiest queries in the right-most cluster of columns, the middle half is represented in two middle clusters (labeled “25%-50%” and “50%-75%”).

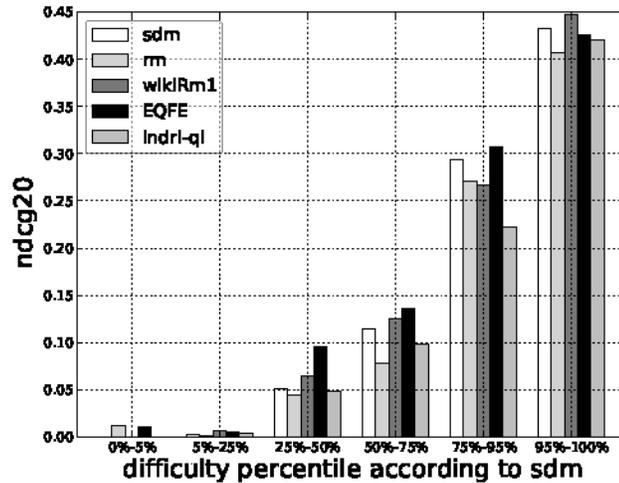
This analysis demonstrates that EQFE especially improves hard queries, for Robust04 and ClueWeb12B all but the top 5% (cf. 7.7a and 7.7c). For ClueWeb09B all queries in the difficult, bottom half (cf. 7.7b) are improved. We want to point out that we achieve this result despite having on average 7 unjudged documents in the top 20 and 2.5 unjudged documents in the top 10 (in both the “5%-25%” and “25%-50%” cluster), which are counted as negatives in the analysis.



(a) Robust04



(b) ClueWeb09B



(c) ClueWeb12B

Figure 7.7: Mean retrieval effectiveness across different query-difficulties, measured according to the percentile of the SDM method. (The hardest queries are on the left)

Dataset	# Queries	# With Freebase	# With Named Entity	# With PER/ORG/LOC
Robust04	249	243	85	49
Clueweb09	200	191	108	80
ClueWeb12	50	48	26	16

Figure 7.8: Number of queries containing different classes of entities (manual labeling)

Dataset	% With Freebase	% With Named Entity	% With PER/ORG/LOC
Robust04	98%	34%	20%
Clueweb09	95%	54%	40%
ClueWeb12	96%	52%	32%

Figure 7.9: Percentage of queries containing different classes of entities

The wikiRM1 method, which is the closest method in spirit to EQFE, demonstrates the opposite characteristic, outperforming EQFE only on “easiest” percentiles.

7.5.2 Entity Analysis of queries

We first examine the characteristics of entities in the queries of these different datasets. For this analysis, the queries were manually labeled for the presence of various classes of entities. We look at three different classes of entities. The first is the most general, whether the entity occurs in Freebase. This includes general concepts. The second class of entities is Named Entities. These are entities that would be tagged by a typical entity recognition system. It includes people, organizations, locations, and other miscellaneous entity types. The last class of entities focuses only on the people, organization, and location entity types. For this analysis, we merely determine whether or not an entity appears anywhere in the query. We do not examine the number of entities in the query or the centrality of the entity to the query.

The entity occurrence statistics for the queries is shown in Tables 7.8 and 7.9. First, we observe that between 95% and 98% of the queries contain at least one mention of a Freebase entity. Many of the entities in the queries are general concepts, such as ‘mammals’, ‘birth rates’, ‘organized crime’, ‘dentistry’, etc... For the web queries, approximately half

Method	Overall	With Freebase	With Named Entity	With PER/ORG/LOC
SDM	26.15	26.61	31.11	27.72
Multiple Source Exp.	30.49	31.02	36.45	31.98
EQFE	32.77	33.33	38.28	33.31

Figure 7.10: Mean Average Precision over different classes of entity queries on Robust04

the queries (54% and 52%) contain a named entity. A smaller percentage of queries for Robust04 contain named entities, only 34%. One reason for this is that web queries are more likely to contain brand names, actors, songs, and movies. Examples of these include ‘Ron Howard,’ ‘I will survive’, ‘Nicolas Cage’, ‘atari’, ‘discovery channel’, ‘espn’, and ‘brooks brothers’.

When the entities are restricted to people, organizations, and locations the fraction of queries containing entities decreases. The fraction of entities that fall into this limited class is between 59% and 74% of the queries containing named entities overall. The entities that are named entities but are not included in this class belong to the “MISC” category and include diseases, songs, movies, naval vessels, drugs, nationalities, buildings, names of government projects, products, treaties, monetary currencies, and others.

7.5.3 Effectiveness by type on Robust04

In this section we describe an analysis of the effectiveness of the previously described classes of queries for the Robust04 dataset. We examine three retrieval models: sequential dependence, multiple source expansion, and entity-based feature expansion. The results are shown in Table 7.10.

We observe that the EQFE expansion model is the best performing model across all types of queries. We also note that queries with entities perform better than those that do not contain them. The gains of queries with Freebase entities are small, which is unsurprising because most of the queries contain at least one entity. However, the entity may not be central to the information need.

The most interesting finding is the comparison of queries with named entities. Queries containing named entities, but not restricted to PER/ORG/LOC shows a large difference over the other classes of queries. This demonstrates that the queries with ‘MISC’ entities perform better than other classes of entity queries. The gains are the largest for this class of queries for EQFE compared with the baseline SDM retrieval model.

7.5.4 Feature-by-Feature Study

We study the contribution of each of the features by ranking the documents according to the feature score and measuring the retrieval effectiveness in MAP. The results for each collection are shown in Figures 7.11, 7.12, and 7.13. The figures show a subset of the top expansion features. The label on the x-axis has three attributes of the entity expansion features: the vocabulary type, feedback source, and number of expansion terms. The vocabulary types are (*A*, *E*, *C*, *W*, and *T* from Figure 7.4). The source is the original query (*Q*), query annotation (*ann*), corpus feedback (*rm*), knowledge base feedback (*kb*), and entity context model feedback (*ecm*). The last usually indicates the number of feedback terms (1, 5, 10, 20, and 50). For *ecm* it indicates the size of the context model window. We note that for several classes of features there are duplicates. These are variations of the same expansion features (for example, top-1 entity, or top-1 non-nil entity).

For the Robust04 collection, the results in Figure 7.11 show that the query and feedback words are the two most effective features. The next set of features shows that entities from the feedback documents perform well, including the entity aliases.

The results for ClueWeb09B in Figure 7.12 are less informative because the combined expansion feature model does not outperform the baseline models when used on the evaluation set. The graph does show that entities from the entity context models appears to have potential to improve over the original query words. The reasons why this feature does not help in the model when combined with the original query words is unclear. This may be due to high variance in the evaluation queries.

For ClueWeb12B in Figure 7.13 the original query is the most effective feature. The entities and words from top-ranked knowledge base articles are also effective. After these features, the entities and entity names from the context models and feedback documents perform well. The words in the feedback documents do not perform as well for this dataset. We hypothesize that this is because web documents have significantly noisier words than newswire collections.

For both ClueWeb collections, the entity context model with window size 8 is the most effective entity context model and a strong feature overall. We also observe that entity aliases are many of the top performing entity features on the web collections. We believe that the entity names are useful because they match documents even when the linked entities may not be found in the document. We believe this is because the entity identifiers in the FACC1 data are highly precise, but may lack recall. Here the name aliases bridge the vocabulary gap between words and entity occurrences.

We note that certain vocabularies such as categories and types do not perform well on their own, but may be helpful in combination with the other features.

7.5.5 Error Analysis of ClueWeb09

We now perform an analysis of the ClueWeb09 results to better understand why EQFE using entity feature expansion does not significantly outperform the baselines.

We first examine the FACC1 query annotations. The FACC1 dataset contains only entity annotations in the description query for 94 of the 200 queries. Upon inspecting the annotations, we found the recall could be improved significantly with further manual annotation. These queries were manually re-annotated to provide entity annotations for 191 of 200 queries. Of the remaining queries without entity links several contain entities that are not noteworthy enough to be included in existing public knowledge bases. These entities are “jax chemical company”, “fickle creek farm”, “sit and reach test”, and “universal animal cuts”. The remaining queries without entities are ambiguous, such as “getting organized”

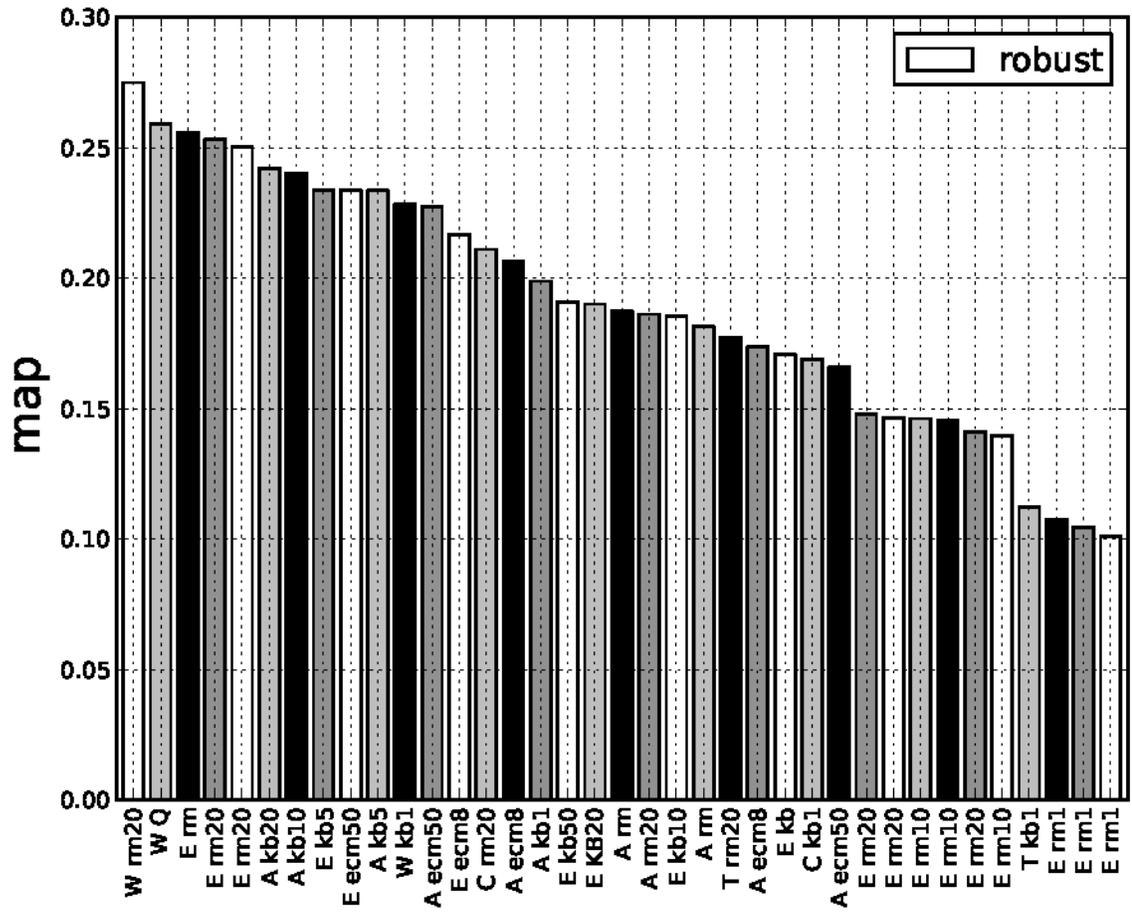


Figure 7.11: Features sorted by retrieval effectiveness on Robust04.

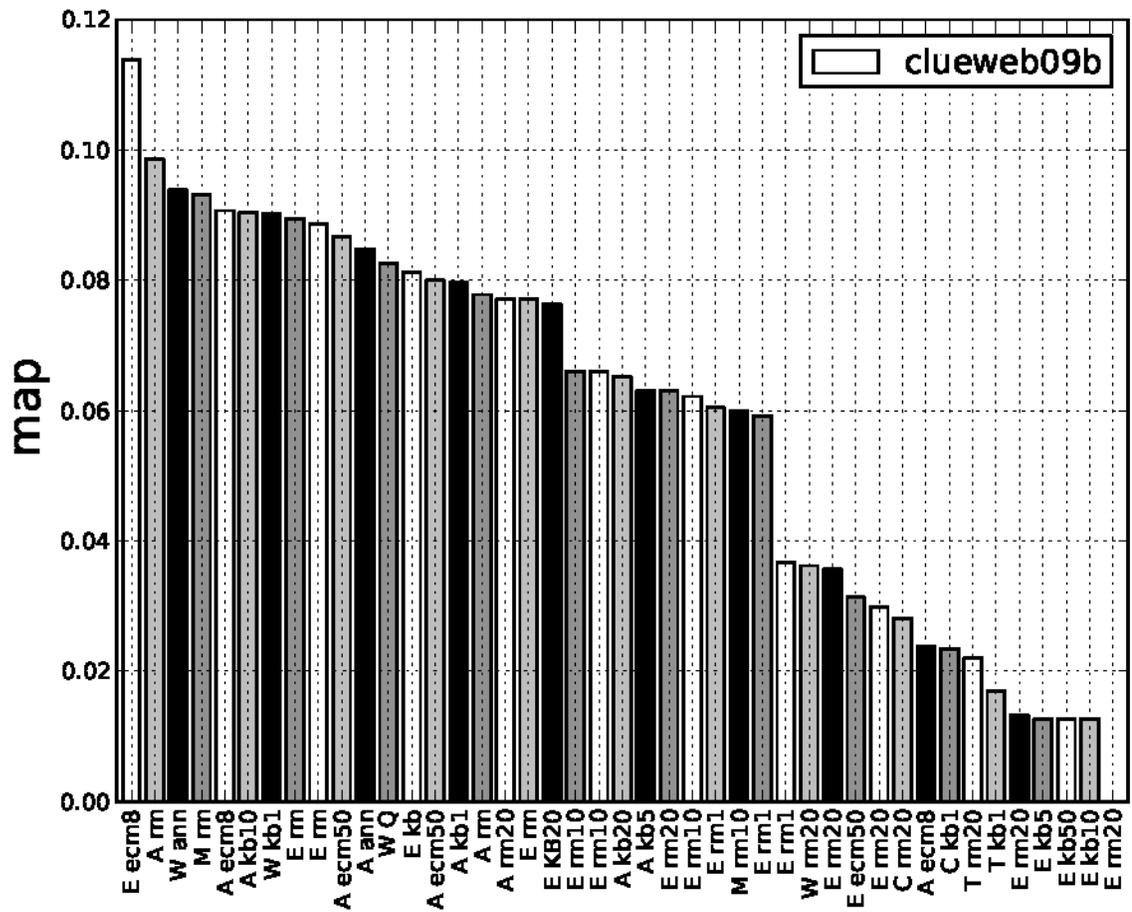


Figure 7.12: Features sorted by retrieval effectiveness on ClueWeb09B.

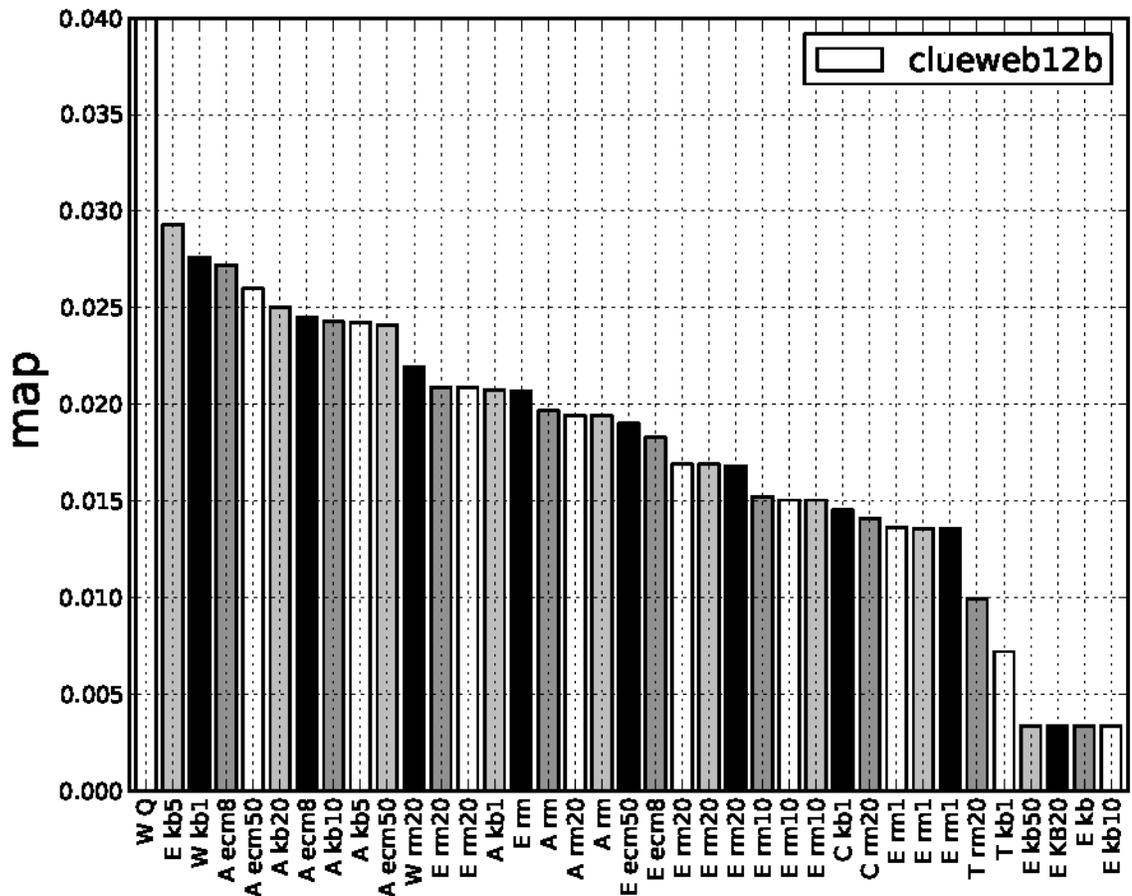


Figure 7.13: Features sorted by retrieval effectiveness on ClueWeb12B.

and “interview thank you”. These are the query entity annotations we report results on in our experiments on ClueWeb09B in this chapter. The revised annotations which will be made publicly available on our website ⁵.

To better understand the role of explicit entities, we use a query model containing only the explicit entity identifiers from revised annotations. This model achieves a MAP score of 0.048, an NDCG@20 of 0.162, and an ERR@20 of 0.123. This is less than half the effectiveness of the SDM baseline. We observe that 72.5% of the documents returned by this model are unjudged. The reasons for this are unclear. Searching using only entity links is a new retrieval paradigm for web search, and it clearly returns results very different from previous text retrieval models. This may indicate that further assessment is required to assess the model effectiveness.

Additionally, we analyze the potential recall on the set of judged relevant documents according to the relevance assessments. We find that 37.4% of the positive judgments in ClueWeb09B do not contain entity links through the FACC1 annotations. The majority of the documents without entity annotations are from Wikipedia, accounting for 24.6% of the documents. Inspecting the FACC1 annotations, only a small subset of Wikipedia articles are included. Further, of the relevant documents that contain at least one entity annotation, we find that 43% of these contain at least one reference to an explicit query annotation. This indicates that there remain significant query-to-document entity recall gaps. This could be the result of missing entity links, or fundamental query-document mismatch.

7.6 Summary

In this chapter we use linked entity representations to perform ad hoc document retrieval. We enrich the query representation with entity features. We present the first known experimental results using entity linked documents and queries for ad hoc document retrieval

⁵<http://ciir.cs.umass.edu/~jdalton/eqfe>

(Contribution 5). Compared with competitive query expansion baselines there is an improvement of 16.4% and 11.5% in MAP on Robust04 and a 14.1% and 32.8% improvement in NDCG@20 for ClueWeb12 (Contribution 2. For ClueWeb09, where results do not significantly improve, we perform an error analysis and identify several important underlying causes for this behavior. In qualitatively different collections (Robust04 and ClueWeb12B), the entity expansion method was on average the strongest performer compared to several state-of-the-art word-based expansion models.

We introduced an extension to existing word dependency models to include dependencies from entity-based features (Contribution 4. One of the key features of this extension is modeling the uncertainty in the entity linking, include the confidence in the links between entity mentions and entities in the knowledge base.

We also defined a new query-specific entity context model that models the feature context of disambiguated entities using retrieved documents (Contribution 3. We found that features from these context models were effective, particularly for web retrieval. We experimented with different sizes and scopes of context and found that a scope of token size eight was effective on web data. The results show that these local context models provide an effective mechanism for identifying the relevance of entities and as a source of expansion features.

CHAPTER 8

CONCLUSION

In this work, we investigated the problem of representing documents using entities, building up increasingly rich entity representations of documents. One of the unique properties of entities is that they represent *things* in the world. One of the fundamental issues in both information extraction and retrieval is that the models make local independence assumptions. The result is incorrect and inconsistent labeling in extraction, and sub-optimal ranking. In this work, we introduced a framework for expanding the local representation with features from entities across documents. We used the enrichment framework to enrich both query and document representation with features from entity mentions.

In Chapter 4 we introduced a new framework for entity-based enrichment. Entity-based enrichment is a focused type of feature expansion where features from similar entity mentions across the collection are used to expand the local representation of a target observation.

We studied the application of entity-based enrichment to three extraction and retrieval tasks: 1) Named entity recognition, 2) Entity linking, and 3) Ad hoc document retrieval. These tasks build upon one another in levels of understanding documents through entities. The first task detects entity mentions, the second links entity mentions to external knowledge resources, and finally the third leverages the disambiguated mentions to improve retrieval effectiveness. We demonstrated how task-specific entity features are used for each of these tasks.

In Chapter 5, we applied the enrichment framework to the task of named entity recognition (NER). The enrichment framework we proposed introduces long-range cross-document

dependencies between similar observations. It uses weighted feature copying from mentions in topically similar passages. In addition to showing that enrichment achieves statistically significant improvements on in-domain accuracy, we show it results in a more robust entity detection model, significantly surpassing other methods when evaluated on out-of-domain data. The enrichment framework allows us to leverage large external sources of unlabeled data. The results show a 6.8% error reduction on newswire and a 19.9% error reduction on out-of-domain book data for named entity recognition.

In Chapter 6, we proposed a method for performing the task of linking detected entity mentions to an external knowledge base using information retrieval. We introduced the *Neighborhood Relevance Model*, which focuses on identifying salient associations between a given entity mention and other entities in the local document neighborhood. The neighborhood relevance model uses the pattern of entity mentions in similar documents to identify salient entity context for disambiguation.

Our experiments on the TAC KBP entity linking data show that this enrichment model outperforms other context weighting models. The results show up to a 16.4% improvement in mean reciprocal rank over local models for entity linking. When combined with a learning-to-rank model that incorporates more text similarity features, the results beat the current best performing systems on in-KB ranking accuracy. Combined with a simple NIL handling strategy the overall effectiveness on all mentions is comparable to, and sometimes better than, other state-of-the-art entity linking systems. We also introduced an enrichment model that used Urban Dictionary to expand the representation of entity mentions in informal forum data.

In Chapter 7, we use linked entity representations to perform ad hoc document retrieval. We enrich the query representation with entity features. We present the first known experimental results using entity linked documents and queries for ad hoc document retrieval. Compared with competitive query expansion baselines (the sequential dependence model with relevance modeling expansion on Wikipedia and the collection) there is an improve-

ment of 16.4% and 11.5% in MAP on Robust04 and a 14.1% and 32.8% improvement in NDCG@20 for ClueWeb12. For ClueWeb09, where results do not significantly improve, we perform an error analysis and identify several important underlying causes for this behavior. In qualitatively different collections (Robust04 and ClueWeb12B), the entity expansion method was on average the strongest performer compared to several leading word-based expansion models.

We also introduced an extension to existing word dependency models to include dependencies from entity-based features (Contribution 4). One of the key features of this extension is modeling the uncertainty in the entity linking, including the confidence in the links between mentions and entities in the knowledge base.

Lastly, we defined a new query-specific entity context model that models the feature context of disambiguated entities using retrieved documents (Contribution 3). We found that features from these context models were effective, particularly for web retrieval. We experimented with different sizes and scopes of context and found that a scope of token size eight was very effective on web data. The results show that these local context models provide an effective mechanism for identifying the relevance of entities and as a source of expansion features.

8.1 Future Work

In our opinion, the use of entities in retrieval is a fledgling area which will continue to grow as search applications become more complex. Entities provide fine-grained conceptual relationships that are shared across documents and provide links to structured attributes in external knowledge sources. Large-scale knowledge bases such as Freebase and Wikipedia are still in their early stages. The use of these and similar large-scale knowledge bases of entities will likely increase as methods for automatic knowledge base construction improve. Currently, these knowledge resources still contain significant gaps which need improvement (an area I intend to pursue in future work). The expressiveness of knowledge bases and the

facts they contain about entities will evolve to better model uncertainty, temporal change, subjective opinions, and complex events.

In Chapter 5, we used enrichment to improve entity detection and classification effectiveness. There are several important extensions to this work. The first is a better method for determining which tokens in the observation sequence require feature enrichment. Using our current heuristics there are over 8,000 queries needed on the small CoNLL test set. The result of improved triggering is that retrieval time could be significantly reduced and the overall effectiveness improved. For sequences with strong evidence feature enrichment is unnecessary and may even degrade effectiveness. A second area is a more principled approach to selecting the source collection to use for enrichment. We would like to utilize strong local evidence within the document and back off to models of similar documents, and finally the entire collection. One possibility would be to investigate techniques similar to the Mixture of Relevance Models (MoRM) (DIAZ and METZLER 2006) and use measures of “concept density” to find rich sub-collections for enrichment.

We focused on enrichment for tagging entities in documents. A similar enrichment approach could be used to identify features for tagging entities in queries. To be useful for retrieval, the types of entities detected needs to evolve significantly. It requires moving beyond traditional named entity types to include other entity classes of interest to users. These include entities in entertainment (such as books, movies, TV shows, recipes), products (including cars, software, phones, and video games), events (concerts, festivals, holidays), and many others.

In addition, other areas of entity detection need to be addressed that impact their utility for retrieval and other applications. For example, entities that are nested or closely related ([Gerald of Wales], [George Washington Bridge], [Kobe Bryant’s wife], [Mrs. George Washington]). Current systems may only detect [Kobe Bryant] and [George Washington], completely missing the implicit reference to another completely different entity. The result is that downstream entity linking systems may not have sufficient context to correctly disam-

biguate these mentions. The result is that important gaps in understanding the document or query remain.

In Chapter 6, we applied enrichment to the task of entity linking. We used enrichment to address the important task of identifying the the salience of disambiguating context. One area for future work is to leverage the enrichment framework to improve linking in other ways. For example, current state-of-the-art models perform “collective” labeling where groups of related mentions are jointly disambiguated to maximize coherence. The enrichment framework in this work could be used to identify groups of related mentions across documents.

One current limitation of existing entity linking is that the task (and data sets) only define one true correct entity as the linking target. The mentioned entity must be the exact entity. For retrieval and other applications alternative task definitions could be useful. For example, linking the mention of a particular room or exhibit in a museum to the museum itself may have significant value, for example geo-locating a particular reference. Another important area for future work is better methods for addressing ‘NIL’ clustering of entities. This area is referred to as cross-document coreference resolution (GOOI and ALLAN 2004; BENJELLOUN *et al.* 2009; WICK *et al.* 2012) and is an area that continues to receive significant attention. The entity enrichment framework we describe could provide a query-focused mechanism of identifying overlapping canopies of related entity mentions for this task.

In Chapter 7, we used the entity representation and cross-document evidence to enrich the query representation. One area for future work is to better understand the types of entities and (and features of those entities) that are most important for retrieval. We believe there are more significant improvements possible by better understanding the types of entities useful for retrieval. One hypothesis is that focusing on common abstract entities such as ‘poverty’, or ‘term limits’ in addition to named entities may be helpful in establishing shared topical context.

One significant area for further analysis is examining the individual components of entity expansion that improve over text features. We examined individual feature-by-feature performance, but we did not explore combinations of features. Future work should pick apart the interactions between the difference classes of expansion features. It is important to understand which attributes of the knowledge base are the most helpful when compared with text alone.

In this initial work we focus on enriching the query as a whole. However, for complex queries with multiple entities, or queries with entities and non-entity words, it is important to model the relationship between these components in the enrichment process. Lastly, in these experiments we focus primarily on entities linkable to knowledge bases. This is because the FACC1 annotations include only these entity mentions. However, we observe that even TREC queries contain tail entities that do not exist in Freebase, like ‘fickle creek farms’. New annotated datasets are needed that include mentions of these entities.

We focused primarily on enriching the query representation. An important area for future work is enriching the document representation. In preliminary work we found that query representations beyond a few short keywords were needed to take advantage of entity-based document representations. Now that we have explored improved query representations, we believe that using the enrichment framework to revisit document representations could be a fruitful area.

Overall, one area for future work is improved modeling of retrieval with other tasks in natural language processing and information extraction. We used the enrichment framework to improve several extraction tasks, and others could similarly be modeled to take advantage of cross-document evidence. These areas include relation extraction, dependency parsing, event extraction, and other tasks in knowledge base construction.

Recently, an emerging trend has been ‘joint inference’ (POON and DOMINGOS 2007; MCCALLUM 2009; WICK *et al.* 2008; WICK *et al.* 2013) of tasks instead of pipelined models. In pipelines errors cascade and there may be a fundamental mismatch between

individual task effectiveness and the final target objective. This has often been the case when information extraction is combined with retrieval, leading to limited improvements. For example, in previous work we demonstrated that different types of coreference errors may disproportionately affect retrieval effectiveness (DALTON *et al.* 2011). An open area for future work would be to extend joint models to include retrieval with other downstream tasks.

In this work, we focus on building up entity-based representation because they offer clear semantics and structured knowledge resources. However, for new domains and obscure information needs, the relevant entity detectors, types, and relations may not exist. An area we propose for future work is query-specific information extraction and knowledge-base construction. This would allow the re-use and extension of existing knowledge along with the construction of new elements ‘on-demand’ in response to an information need. We recently proposed one step in this direction with the *knowledge sketch*, (DALTON and DIETZ 2013b) which assembles relevant fragments of knowledge from existing knowledge bases as well as constructs new data extracted from retrieved documents to create a query-specific knowledge artifact. A knowledge sketch contains distributions over relevant entities, documents, and relationships between entities.

We have focused on building up increasingly complex entity-based representations of documents and queries. This is an active area of research which we believe will continue to grow in importance as search applications evolve in the future.

BIBLIOGRAPHY

- ATTAR, R. and A. S. FRAENKEL, 1977 Local feedback in full-text retrieval systems. *Journal of the ACM (JACM)* 24(3): 397–417.
- AUER, S., C. BIZER, G. KOBILAROV, J. LEHMANN, R. CYGANIAK, and Z. IVES, 2007 DBpedia: a nucleus for a web of open data. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, ISWC'07/ASWC'07*, Berlin, Heidelberg, pp. 722–735. Springer-Verlag.
- BAGGA, A. and B. BALDWIN, 1998 Algorithms for Scoring Coreference Chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pp. 563–566.
- BALOG, K., M. BRON, and M. DE RIJKE, 2011 Query modeling for entity search based on terms, categories, and examples. *ACM Transactions on Information Systems (TOIS)* 29(4): 22.
- BALOG, K., D. CARMEL, A. P. DE VRIES, D. M. HERZIG, P. MIKA, H. ROITMAN, R. SCHENKEL, P. SERDYUKOV, and T. T. DUC, 2012 The first joint international workshop on entity-oriented and semantic search (JIWES). In *ACM SIGIR Forum*, Volume 46, pp. 87–94. ACM.
- BALOG, K. and M. DE RIJKE, 2008 Non-local Evidence for Expert Finding. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, New York, NY, USA, pp. 489–498. ACM.
- BALOG, K., P. SERDYUKOV, and A. P. DE VRIES, 2011 Overview of the TREC 2011 Entity Track. In *Proceedings of the Text REtrieval Conference (TREC)*.

- BALOG, K., P. THOMAS, N. CRASWELL, I. SOBOROFF, P. BAILEY, and A. P. VRIES, 2008 Overview of the TREC 2008 Enterprise Track.
- BENDERSKY, M. and W. B. CROFT, 2012 Modeling Higher-Order Term Dependencies in Information Retrieval using Query Hypergraphs. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 941–950.
- BENDERSKY, M., W. B. CROFT, and D. A. SMITH, 2010 Structural annotation of search queries using pseudo-relevance feedback. In *Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10*, New York, NY, USA, pp. 1537–1540. ACM.
- BENDERSKY, M., D. METZLER, and W. B. CROFT, 2012 Effective query formulation with multiple information sources. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12*, New York, NY, USA, pp. 443–452. ACM.
- BENJELLOUN, O., H. G. MOLINA, D. MENESTRINA, Q. SU, S. E. WHANG, and J. WIDOM, 2009 Swoosh: a generic approach to entity resolution. *The VLDB Journal* 18(1): 255–276.
- BENNETT, P., E. GABRILOVICH, J. KAMPS, and J. KARLGREN, 2013 Sixth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'13). In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, CIKM '13*, New York, NY, USA, pp. 2543–2544. ACM.
- BERNERS-LEE, T., J. HENDLER, O. LASSILA, and OTHERS, 2001 The semantic web. *Scientific american* 284(5): 28–37.

- BLAIR, D. C., 1979 *Information Retrieval*, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: \$32.50. *Journal of the American Society for Information Science* 30(6): 374–375.
- BLANCO, R., H. HALPIN, D. M. HERZIG, P. MIKA, J. POUND, H. S. THOMPSON, and T. T. DUC, 2011 Entity search evaluation over structured web data. In *Proceedings of the 1st international workshop on entity-oriented search workshop (SIGIR 2011)*, ACM, New York.
- BLEI, D. M., A. Y. NG, and M. I. JORDAN, 2003 Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3: 993–1022.
- BOLLACKER, K., C. EVANS, P. PARITOSH, T. STURGE, and J. TAYLOR, 2008 Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, New York, NY, USA, pp. 1247–1250. ACM.
- BROWN, P. F., P. V. DESOUZA, R. L. MERCER, V. J. D. PIETRA, and J. C. LAI, 1992 Class-based n-gram models of natural language. *Computational linguistics* 18(4): 467–479.
- BUCKLEY, C. Automatic Query Expansion Using SMART : TREC 3. In *In Proceedings of The third Text REtrieval Conference (TREC-3)*, pp. 69–80.
- BUNESCU, R. and R. J. MOONEY, 2004 Collective information extraction with relational Markov networks. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.
- BUNESCU, R. and M. PASCA, 2006 Using Encyclopedic Knowledge for Named Entity Disambiguation. In *European Chapter of the Association for Computational Linguistics (EACL-06)*, pp. 9–16.

- CALLAN, J. P., W. B. CROFT, and J. BROGLIO, 1994 TREC and TIPSTER Experiments with INQUERY. In *Information Processing & Management*, pp. 31–3. Morgan Kaufmann.
- CHAPELLE, O., D. METLZER, Y. ZHANG, and P. GRINSPAN, 2009 Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09*, New York, NY, USA, pp. 621–630. ACM.
- CHU-CARROLL, J., J. PRAGER, C. WELTY, K. CZUBA, and D. FERRUCCI, 2003 A Multi-Strategy and Multi-Source Approach to Question Answering. In *Proceedings of Text REtrieval Conference*.
- CLEVERDON, C. W., 1991 The Significance of the Cranfield Tests on Index Languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '91*, New York, NY, USA, pp. 3–12. ACM.
- CONNELL, M., A. FENG, G. KUMARAN, H. RAGHAVAN, C. SHAH, and J. ALLAN, 2004 UMass at TDT 2004. In *Topic Detection and Tracking Workshop Report*.
- CONRAD, J. G. and M. H. UTT, 1994 A system for discovering relationships by feature extraction from text databases. In *SIGIR'94*, pp. 260–270. Springer.
- CORMACK, G. V., M. D. SMUCKER, and C. L. A. CLARKE, 2011 Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval* 14(5): 441–465.
- CRASWELL, N., A. P. DE VRIES, and I. SOBOROFF, 2005 Overview of the TREC 2005 Enterprise Track. In *TREC*.
- CROFT, B. W. and D. J. HARPER, 1979 Using Probabilistic Models of Document Retrieval Without Relevance Information. *Journal of Documentation* 35(4): 285–295.

- CROFT, W. B., T. J. LUCIA, J. CRINGEAN, and P. WILLETT, 1989 Retrieving documents by plausible inference: An experimental study. *Information Processing & Management* 25(6): 599–614.
- CUCERZAN, S., 2007 Large-Scale Named Entity Disambiguation Based on Wikipedia Data. EMNLP.
- CUCERZAN, S., 2011 TAC Entity Linking by Performing Full-document Entity Extraction and Disambiguation. In *Proceedings of the Text Analysis Conference 2011*.
- DALTON, J., J. ALLAN, and D. A. SMITH, 2011 Passage retrieval for incorporating global evidence in sequence labeling. In *Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11*, New York, NY, USA, pp. 355–364. ACM.
- DALTON, J., R. BLANCO, and P. MIKA, 2011 Coreference Aware Web Object Retrieval. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11*, New York, NY, USA, pp. 211–220. ACM.
- DALTON, J. and L. DIETZ, 2012 Bi-directional Linkability From Wikipedia to Documents and Back Again: UMass at TREC 2012 Knowledge Base Acceleration Track. In *Proceedings of the Text REtrieval Conference (TREC)*.
- DALTON, J. and L. DIETZ, 2013a A Neighborhood Relevance Model for Entity Linking. In *Proceedings of the 10th International Conference in the RIAO series (OAIR)*, RIAO '13, New York, NY, USA. ACM.
- DALTON, J. and L. DIETZ, 2013b Constructing Query-specific Knowledge Bases. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC '13*, New York, NY, USA, pp. 55–60. ACM.
- DALTON, J. and L. DIETZ, 2013c UMass at TAC KBP 2013 Entity Linking: Query Expansion using Urban Dictionary. In *Proceedings of the Text Analysis Conference (TAC KBP)*.

- DALTON, J., L. DIETZ, and J. ALLAN, 2014 Entity Query Feature Expansion using Knowledge Base Links. In *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 14, Gold Coast, Australia. ACM.
- DALTON, J. and S. HUSTON, 2010 Semantic entity retrieval using web queries over structured RDF data. In *Proc. of the 3rd Intl. Semantic Search Workshop*.
- DEMARTINI, G., T. IOFCIU, and A. VRIES, 2010 Overview of the INEX 2009 Entity Ranking Track. In S. Geva, J. Kamps, and A. Trotman (Eds.), *Focused Retrieval and Evaluation*, Volume 6203 of *Lecture Notes in Computer Science*, pp. 254–264. Springer Berlin Heidelberg.
- DIAZ, F., 2008 Autocorrelation and Regularization of Query-based Information Retrieval Scores. Ir, University of Massachusetts.
- DIAZ, F. and D. METZLER, 2006 Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, New York, NY, USA, pp. 154–161. ACM.
- DUMAIS, S. T., 1995 Latent Semantic Indexing (LSI): TREC-3 Report. In *Overview of the Third Text REtrieval Conference*, pp. 219–230.
- EFRON, M., P. ORGANISCIAK, and K. FENLON, 2012 Improving Retrieval of Short Texts Through Document Expansion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, New York, NY, USA, pp. 911–920. ACM.
- EGOZI, O., E. GABRILOVICH, and S. MARKOVITCH, 2008 Concept-based feature generation and selection for information retrieval. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, AAAI'08, pp. 1132–1137. AAAI Press.

- ELLIS, J., X. LI, K. GRIFFITT, S. M. STRASSEL, and J. WRIGHT, 2011 Linguistic resources for 2012 knowledge base population evaluations.
- FERRUCCI, D., E. BROWN, J. CHU-CARROLL, J. FAN, D. GONDEK, A. A. KALYANPUR, A. LALLY, J. W. MURDOCK, E. NYBERG, J. PRAGER, N. SCHLAEFER, and C. WELTY, 2010 Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31(3).
- FINKEL, J. R., T. GRENAGER, and C. MANNING, 2005 Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL*, pp. 363–370.
- FRANK, J. R., M. KLEIMAN-WEINER, D. A. ROBERTS, F. NIU, C. ZHANG, C. RE, and I. SOBOROFF, 2012 Building an Entity-Centric Stream Filtering Test Collection for TREC 2012. In *Proceedings of the Text REtrieval Conference (TREC)*.
- FURNAS, G. W., T. K. LANDAUER, L. M. GOMEZ, and S. T. DUMAIS, 1987 The Vocabulary Problem in Human-system Communication. *Commun. ACM* 30(11): 964–971.
- GABRILOVICH, E. and S. MARKOVITCH, 2006 Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2, AAAI'06*, pp. 1301–1306. AAAI Press.
- GABRILOVICH, E. and S. MARKOVITCH, 2007 Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artificial intelligence, IJCAI'07*, San Francisco, CA, USA, pp. 1606–1611. Morgan Kaufmann Publishers Inc.
- GABRILOVICH, E., M. RINGGAARD, and A. SUBRAMANYA, 2013 FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0).

- GAUCH, S., J. CHAFFEE, and A. PRETSCHNER, 2003 Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys. 1(3-4)*: 219–234.
- GOLDER, S. A. and B. A. HUBERMAN, 2006 Usage patterns of collaborative tagging systems. *Journal of Information Science 32(2)*: 198–208.
- GOOI, C. H. and J. ALLAN, 2004 Cross-Document Coreference on a Large Scale Corpus. In Daniel and S. Roukos (Eds.), *HLT-NAACL*, Boston, Massachusetts, USA, pp. 9–16. Association for Computational Linguistics.
- GOTTIPATI, S. and J. JIANG, 2011 Linking entities to a knowledge base with query expansion. In *EMNLP, EMNLP '11*, Stroudsburg, PA, USA, pp. 804–813. Association for Computational Linguistics.
- GUO, J., G. XU, X. CHENG, and H. LI, 2009 Named Entity Recognition in Query. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09*, New York, NY, USA, pp. 267–274. ACM.
- HOFFART, J., M. A. YOSEF, I. BORDINO, H. FÜRSTENAU, M. PINKAL, M. SPAN-
IOL, B. TANEVA, S. THATER, and G. WEIKUM, 2011 Robust disambiguation of named entities in text. In *EMNLP, EMNLP '11*, Stroudsburg, PA, USA, pp. 782–792. Association for Computational Linguistics.
- HOFMANN, T., 2001 Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning 42(1)*.
- HUANG, D. W., Y. XU, A. TROTMAN, and S. GEVA, 2008 Focused Access to XML Documents. Chapter Overview of INEX 2007 Link the Wiki Track, pp. 373–387. Berlin, Heidelberg: Springer-Verlag.
- JALEEL, N. A., J. ALLAN, W. B. CROFT, F. DIAZ, L. LARKEY, X. LI, M. SMUCKER, and C. WADE, 2005 UMass at TREC 2004: Novelty and HARD. Proc. TREC 2004, <http://trec.nist.gov/>.

- JI, H. and R. GRISHMAN, 2011 Knowledge base population: successful approaches and challenges. In *HLT, HLT '11*, Stroudsburg, PA, USA, pp. 1148–1158. Association for Computational Linguistics.
- JI, H., R. GRISHMAN, and H. DANG, 2011 Overview of the TAC2011 Knowledge Base Population Track. In *Text Analysis Conference*.
- JING, Y. and CROFT, 1994 An Association Thesaurus for Information Retrieval. Technical report, Amherst, MA, USA.
- JONES, K. S. and E. O. BARBER, 1971 What makes an automatic keyword classification effective? *Journal of the American Society for Information Science* 22(3): 166–175.
- JOYCE, T. and R. M. NEEDHAM, 1958 The thesaurus approach to information retrieval. *American Documentation* 9(3): 192–197.
- KAMPS, J., J. KARLGREN, P. MIKA, and V. MURDOCK, 2012 Fifth Workshop on Exploiting Semantic Annotations in Information Retrieval: ESAIR'12). In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, New York, NY, USA, pp. 2772–2773. ACM.
- KAMPS, J., J. KARLGREN, and R. SCHENKEL, 2011 Report on the Third Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR). *SIGIR Forum* 45(1): 33–41.
- KANANI, P. H. and A. K. MCCALLUM, 2012 Selecting actions for resource-bounded information extraction using reinforcement learning. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12*, New York, NY, USA, pp. 253–262. ACM.
- KAPTEIN, R., P. SERDYUKOV, and J. KAMPS, 2010 Linking wikipedia to the web. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10*, New York, NY, USA, pp. 839–840. ACM.

- KIM, T., E. F. SANG, and F. DE MEULDER, 2003 Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *CoNLL*, pp. 142–147.
- KOLLER, D. and N. FRIEDMAN, 2009 *Probabilistic graphical models: principles and techniques*. MIT press.
- KRISHNAN, V. and C. D. MANNING, 2006 An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. In *ACL*, pp. 1121–1128.
- KROVETZ, R., 1993 Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 191–202. ACM.
- KROVETZ, R. and W. B. CROFT, 1989 Word Sense Disambiguation Using Machine-readable Dictionaries. In *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '89*, New York, NY, USA, pp. 127–136. ACM.
- KROVETZ, R. and W. B. CROFT, 1992 Lexical Ambiguity and Information Retrieval. *ACM Trans. Inf. Syst.* 10(2): 115–141.
- KULKARNI, S., A. SINGH, G. RAMAKRISHNAN, and S. CHAKRABARTI, 2009 Collective annotation of Wikipedia entities in web text. In *KDD, KDD '09*, New York, NY, USA, pp. 457–466. ACM.
- KUMARAN, G. and J. ALLAN, 2004 Text Classification and Named Entities for New Event Detection. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04*, New York, NY, USA, pp. 297–304. ACM.
- LAFFERTY, J., A. MCCALLUM, and F. PEREIRA, 2001 Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pp. 282–289.

- LANG, H., D. METZLER, B. WANG, and J. T. LI, 2010 Improved latent concept expansion using hierarchical markov random fields. In *CIKM*.
- LAVRENKO, V. and W. B. CROFT, 2001 Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, New York, NY, USA, pp. 120–127. ACM.
- LEHMANN, J., S. MONAHAN, L. NEZDA, A. JUNG, and Y. SHI, 2010 LCC Approaches to Knowledge Base Population at TAC 2010. Technical report.
- LENAT, D. B., 1995 CYC: A Large-scale Investment in Knowledge Infrastructure. *Commun. ACM* 38(11): 33–38.
- LEWIS, D. D., Y. YANG, T. G. ROSE, and F. LI, 2004 RCV1: A New Benchmark Collection for Text Categorization Research. *J. Mach. Learn. Res.* 5: 361–397.
- LI, X., Y.-Y. WANG, and A. ACERO, 2009 Extracting Structured Information from User Queries with Semi-supervised Conditional Random Fields. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, New York, NY, USA, pp. 572–579. ACM.
- LIN, J. and D. DEMNER-FUSHMAN, 2006 The Role of Knowledge in Conceptual Retrieval: A Study in the Domain of Clinical Medicine. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, New York, NY, USA, pp. 99–106. ACM.
- LIPSCOMB, C. E., 2000 Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88(3): 265.
- LIU, X. and W. B. CROFT, 2004 Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, New York, NY, USA, pp. 186–193. ACM.

- LIU, X., S. ZHANG, F. WEI, and M. ZHOU, 2011 Recognizing Named Entities in Tweets. In *ACL*.
- LUHN, H. P., 1957 A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development* 1(4): 309–317.
- LV, Y. and C. ZHAI, 2010 Positional Relevance Model for Pseudo-relevance Feedback. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10*, New York, NY, USA, pp. 579–586. ACM.
- MCCALLUM, A., 2009 Joint Inference for Natural Language Processing. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL '09*, Stroudsburg, PA, USA, pp. 1. Association for Computational Linguistics.
- MCCALLUM, A., K. SCHULTZ, and S. SINGH, 2009 Factorie: Probabilistic programming via imperatively defined factor graphs. In *In Advances in Neural Information Processing Systems 22*, pp. 1249–1257.
- MCNAMEE, P., V. STOYANOV, J. MAYFIELD, T. FININ, T. OATES, T. XU, D. W. OARD, and D. LAWRIE, 2012 HLTCOE Participation at TAC 2012: Entity Linking and Cold Start Knowledge Base Construction. In *Proceedings of the Text Analysis Conference (TAC KBP)*.
- MEI, Q., D. ZHANG, and C. ZHAI, 2008 A general optimization framework for smoothing language models on graph structures. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08*, New York, NY, USA, pp. 611–618. ACM.
- MEIJ, E., M. BRON, L. HOLLINK, B. HUURNINK, and M. RIJKE, 2009 Learning Semantic Query Suggestions. In *Proceedings of the 8th International Semantic Web Conference, ISWC '09*, Berlin, Heidelberg, pp. 424–440. Springer-Verlag.

- MEIJ, E. and M. DE RIJKE, 2010 Supervised query modeling using Wikipedia. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2010.
- MEIJ, E. J., 2010 *Combining concepts and language models for information access*.
- METZLER, D. and W. B. CROFT, 2005 A Markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, New York, NY, USA, pp. 472–479. ACM.
- METZLER, D. and W. B. CROFT, 2007 Latent concept expansion using markov random fields. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, New York, NY, USA, pp. 311–318. ACM.
- METZLER, D., J. NOVAK, H. CUI, and S. REDDY, 2009 Building Enriched Document Representations Using Aggregated Anchor Text. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, New York, NY, USA, pp. 219–226. ACM.
- MIHALCEA, R. and D. MOLDOVAN, 2000 Semantic Indexing Using WordNet Senses. In *Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11*, RANLPIR '00, Stroudsburg, PA, USA, pp. 35–45. Association for Computational Linguistics.
- MIKA, P. and T. POTTER, 2012 Metadata statistics for a large web corpus. In *Proceedings of the Linked Data Workshop (LDOW) at the International World Wide Web Conference*.
- MIKHEEV, A., 1999 A Knowledge-free Method for Capitalized Word Disambiguation. In *ACL*, pp. 159–166.

- MILLER, G. A., 1995 WordNet: A Lexical Database for English. *Commun. ACM* 38(11): 39–41.
- MILNE, D. and I. H. WITTEN, 2008 Learning to Link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, New York, NY, USA, pp. 509–518. ACM.
- PASCA, M., 2013 Open-Domain Fine-Grained Class Extraction from Web Search Queries. In *EMNLP*, pp. 403–414.
- PETKOVA, D. and W. B. CROFT, 2007 Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07*, New York, NY, USA, pp. 731–740. ACM.
- PINTO, D., A. MCCALLUM, X. WEI, and W. B. CROFT, 2003 Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03*, New York, NY, USA, pp. 235–242. ACM.
- POON, H. and P. DOMINGOS, 2007 Joint Inference in Information Extraction. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 1, AAAI'07*, pp. 913–918. AAAI Press.
- RAGHAVAN, H., J. ALLAN, and A. MCCALLUM, 2004 An exploration of entity models, collective classification and relation description. In *KDD Workshop on Link Analysis and Group Detection*, pp. 1–10.
- RATINOV, L. and D. ROTH, 2009 Design challenges and misconceptions in named entity recognition. In *CoNLL*, pp. 147–155.
- RATINOV, L., D. ROTH, D. DOWNEY, and M. ANDERSON, 2011 Local and global algorithms for disambiguation to wikipedia. In *ACL*.

- ROBERTSON, S. E., 1977 The probability ranking principle in IR. *Journal of documentation* 33(4): 294–304.
- ROCCHIO, J. J., 1971 Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall Series in Automatic Computation, Chapter 14, pp. 313–323. Prentice-Hall, Englewood Cliffs NJ.
- SALTON, G., 1968 *Automatic Information Organization and Retrieval*. McGraw Hill Text.
- SALTON, G., 1972 A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science* 23(2): 75–84.
- SALTON, G. and C. BUCKLEY, 1990 *Improving retrieval performance by relevance feedback*, Volume 41, pp. 288–297. San Francisco, CA, USA: Wiley.
- SANDERSON, M., 1994 Word Sense Disambiguation and Information Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94*, New York, NY, USA, pp. 142–151. Springer-Verlag New York, Inc.
- SCHLAEFER, N., J. C. CARROLL, E. NYBERG, J. FAN, W. ZADROZNY, and D. FER-
RUCCI, 2011 Statistical source expansion for question answering. In *Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11*, New York, NY, USA, pp. 345–354. ACM.
- SCHÜTZE, H. and J. O. PEDERSEN, 1995 Information Retrieval Based on Word Senses.
- SHAH, C., W. B. CROFT, and D. JENSEN, 2006 Representing Documents with Named Entities for Story Link Detection (SLD). In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM '06*, New York, NY, USA, pp. 868–869. ACM.

- SINGHAL, A. and F. PEREIRA, 1999 Document Expansion for Speech Retrieval. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, New York, NY, USA, pp. 34–41. ACM.
- SMUCKER, M. D., J. ALLAN, and B. CARTERETTE, 2007 A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, New York, NY, USA, pp. 623–632. ACM.
- STOYANOV, V., J. MAYFIELD, T. XU, D. W. OARD, D. LAWRIE, T. OATES, and T. FININ, 2012 A context-aware approach to entity linking. In *AKBC-WEKEX*, AKBC-WEKEX '12, Stroudsburg, PA, USA, pp. 62–67. Association for Computational Linguistics.
- SUCHANEK, F. M., S. RIEDEL, S. SINGH, and P. P. TALUKDAR, 2013 AKBC 2013: third workshop on automated knowledge base construction. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 2539–2540. ACM.
- SUTTON, C. and A. MCCALLUM, 2004 Collective segmentation and labeling of distant entities in information extraction. In *ICML Workshop on Statistical Relational Learning and its Connections to Other Fields*.
- SUTTON, C., M. SINDELAR, and A. MCCALLUM, 2006 Reducing weight undertraining in structured discriminative learning. In *HLT-NAACL*, pp. 89–95.
- TAO, T., X. WANG, Q. MEI, and C. ZHAI, 2006a Language Model Information Retrieval with Document Expansion. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, Stroudsburg, PA, USA, pp. 407–414. Association for Computational Linguistics.

- TAO, T., X. WANG, Q. MEI, and C. ZHAI, 2006b Language model information retrieval with document expansion. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, Stroudsburg, PA, USA, pp. 407–414. Association for Computational Linguistics.
- TRAN, T., P. MIKA, H. WANG, and M. GROBELNIK, 2010 SEMSEARCH '10: Proceedings of the 3rd International Semantic Search Workshop. New York, NY, USA. ACM.
- TRAN, T., P. MIKA, H. WANG, and M. GROBELNIK, 2011 SemSearch'11: The 4th Semantic Search Workshop. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, New York, NY, USA, pp. 315–316. ACM.
- VILAIN, M., J. HUGGINS, and B. WELLNER, 2009 A simple feature-copying approach for long-distance dependencies. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, Stroudsburg, PA, USA, pp. 192–200. Association for Computational Linguistics.
- VOORHEES, E. M., 1999 Natural language processing and information retrieval. In *Information Extraction*, pp. 32–48. Springer.
- WANG, Q., J. KAMPS, G. R. CAMPS, M. MARX, A. SCHUTH, M. THEOBALD, S. GURAJADA, and A. MISHRA, 2011 Overview of the INEX 2012 Linked Data Track. In *INitiative for the Evaluation of XML Retrieval (INEX)*.
- WEI, X. and W. B. CROFT, 2006 LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, New York, NY, USA, pp. 178–185. ACM.

- WEI, X. and W. B. CROFT, 2007 Investigating retrieval performance with manually-built topic models. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, Paris, France, France, pp. 333–349. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.
- WELLING, M., M. ROSEN-ZVI, and G. E. HINTON, 2004 Exponential family harmoniums with an application to information retrieval. In *Neural Information Processing Systems (NIPS) 17*.
- WEST, R., E. GABRILOVICH, K. MURPHY, S. SUN, R. GUPTA, and D. LIN, 2014 Knowledge Base Completion via Search-Based Question Answering. In *WWW*.
- WICK, M., S. SINGH, and A. MCCALLUM, 2012 A discriminative hierarchical model for fast coreference at large scale. In *ACL, ACL '12*, Stroudsburg, PA, USA, pp. 379–388. Association for Computational Linguistics.
- WICK, M., S. SINGH, H. PANDYA, and A. MCCALLUM, 2013 A Joint Model for Discovering and Linking Entities. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC '13*, New York, NY, USA, pp. 67–72. ACM.
- WICK, M. L., K. ROHANIMANESH, K. SCHULTZ, and A. MCCALLUM, 2008 A Unified Approach for Schema Matching, Coreference and Canonicalization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, New York, NY, USA, pp. 722–730. ACM.
- XU, J. and W. B. CROFT, 1996 Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 4–11. ACM.
- XU, J. and W. B. CROFT, 2000 Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.* 18(1): 79–112.
- XU, Y., G. J. F. JONES, and B. WANG, 2009 Query Dependent Pseudo-relevance Feedback Based on Wikipedia. In *Proceedings of the 32Nd International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR '09, New York, NY, USA, pp. 59–66. ACM.

YI, X. and J. ALLAN, 2009 A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, Volume 5478 of *ECIR '09*, Berlin, Heidelberg, pp. 29–41. Springer-Verlag.

YI, X. and J. ALLAN, 2010 A Content Based Approach for Discovering Missing Anchor Text for Web Search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10*, New York, NY, USA, pp. 427–434. ACM.

ZHAI, C. and J. LAFFERTY, 2001 Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01*, New York, NY, USA, pp. 403–410. ACM.