Semantic Search on Text and Knowledge Bases

Hannah Bast
University of Freiburg
bast@cs.uni-freiburg.de

Björn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Elmar Haussmann
University of Freiburg
haussmann@cs.uni-freiburg.de

the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

H. Bast, B. Buchhold, E. Haussmann. *Semantic Search on Text and Knowledge Bases*. Foundations and Trends® in Information Retrieval, vol. 10, no. 2-3, pp. 119–271, 2016.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-165-8

Foundations and Trends® in
Information Retrieval
Volume 10, Issue 2-3, 2016

Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2016, Volume 10, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in Information Retrieval
Vol. 10, No. 2-3 (2016) 119–271

DOI: 10.1561/1500000032

the essence of knowledge

Semantic Search on Text and Knowledge Bases

Hannah Bast
University of Freiburg
bast@cs.uni-freiburg.de

Björn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Elmar Haussmann
University of Freiburg
haussmann@cs.uni-freiburg.de

Contents

Abstract

This article provides a comprehensive overview of the broad area of semantic search on text and knowledge bases. In a nutshell, semantic search is "search with meaning". This "meaning" can refer to various parts of the search process: understanding the query (instead of just finding matches of its components in the data), understanding the data (instead of just searching it for such matches), or representing knowledge in a way suitable for meaningful retrieval.

Semantic search is studied in a variety of different communities with a variety of different views of the problem. In this survey, we classify this work according to two dimensions: the type of data (text, knowledge bases, combinations of these) and the kind of search (keyword, structured, natural language). We consider all nine combinations. The focus is on fundamental techniques, concrete systems, and benchmarks. The survey also considers advanced issues: ranking, indexing, ontology matching and merging, and inference. It also provides a succinct overview of natural language processing techniques that are useful for semantic search: POS tagging, named-entity recognition and disambiguation, sentence parsing, and word vectors.

The survey is as self-contained as possible, and should thus also serve as a good tutorial for newcomers to this fascinating and highly topical field.

H. Bast, B. Buchhold, E. Haussmann. *Semantic Search on Text and Knowledge Bases*. Foundations and Trends® in Information Retrieval, vol. 10, no. 2-3, pp. 119–271, 2016.

DOI: 10.1561/1500000032.

1

Introduction

1.1 Motivation for this Survey

This is a survey about the broad field of semantic search. Semantics is the study of meaning.[1] In a nutshell, therefore, it could be said that semantic search is *search with meaning*.

Let us first understand this by looking at the opposite. Only a decade ago, search engines, including the big web search engines, were still mostly *lexical*. By lexical, we here mean that the search engine looks for literal matches of the query words typed by the user or variants of them, without making an effort to understand what the whole query actually means.

Consider the query *university freiburg* issued to a web search engine. Clearly, the homepage of the University of Freiburg is a good match for this query. To identify this page as a match, the search engine does not need to understand what the two query words *university* and *freiburg* actually mean, nor what they mean together. In fact, the university homepage contains these two words in its title (and, as a

[1]The word comes from the ancient greek word *sēmantikós*, which means *important*.

matter of fact, no other except the frequent word *of*). Further, the page is at the top level of its domain, as can be seen from its URL: `http://www.uni-freiburg.de`. Even more, the URL consists of parts of the query words. All these criteria are easy to check, and they alone make this page a very good candidate for the top hit of this query. No deeper understanding of what the query actually "meant" or what the homepage is actually "about" were needed.[2]

Modern search engines go more and more in the direction of accepting a broader variety of queries, actually trying to "understand" them, and providing the most appropriate answer in the most appropriate form, instead of just a list of (excerpts from) matching documents.

For example, consider the two queries *computer scientists* and *female computer scientists working on semantic search*. The first query is short and simple, the second query is longer and more complex. Both are good examples of what we would call semantic search. The following discussion is independent of the exact form of these queries. They could be formulated as keyword queries like above. They could be formulated in the form of complete natural language queries. Or they could be formulated in an abstract query language. The point here is what the queries are asking for.

To a human, the intention of both of these queries is quite clear: the user is (most likely) looking for scientists of a certain kind. Probably a list of them would be nice, with some basic information on each (for instance, a picture and a link to their homepage). For the query *computer scientists*, Wikipedia happens to provide a page with a corresponding list and matching query words.[3] Correspondingly, the list is also contained in DBpedia, a database containing the structured knowledge from Wikipedia. But in both cases it is a manually compiled list, limited to relatively few better-known computer scientists. For the second query (*female computer scientists working on semantic search*), there is no single web page or other document with a corresponding

[2]In this simple example, we are leaving aside the important issue of *spam*. That is, someone deliberately putting misleading keywords in the title or even in the URL, in order to fool search engines, and thus users, to consider the web page relevant. Note that this query could also be solved using clickthrough data; see Section 1.2.2.

[3]`http://en.wikipedia.org/wiki/List_of_computer_scientists`

list, let alone one matching the query words. Given the specificity of the query, it is also unlikely that someone will ever manually compile such a list (in whatever format) and maintain it. Note that both lists are constantly changing over time, since new researchers may join any time.

In fact, even individual web pages matching the query are unlikely to contain most of the query words. A computer scientist does not typically put the words *computer scientist* on his or her homepage. A female computer scientist is unlikely to put the word *female* on her homepage. The homepage probably has a section on that particular scientist's research interests, but this section does not necessarily contain the word *working* (maybe it contains a similar word, or maybe no such word at all, but just a list of topics). The topic *semantic search* will probably be stated on a matching web page, though possibly in a different formulation, for example, *intelligent search* or *knowledge retrieval*.

Both queries are thus good examples, where search needs to go beyond mere lexical matching of query words in order to provide a satisfactory result to the user. Also, both queries (in particular, the second one) require that information from several different sources is brought together to answer the query satisfactorily. Those information sources might be of different kinds: (unstructured) text as well as (structured) knowledge bases.

There is no exact definition of what semantic search is. In fact, semantic search means a lot of different things to different people. And researchers from many different communities are working on a large variety of problems related to semantic search, often without being aware of related work in other communities. *This is the main motivation behind this survey.*

When writing the survey, we had two audiences in mind: (i) newcomers to the field, and (ii) researchers already working on semantic search. Both audiences should get a comprehensive overview of which approaches are currently pursued in which communities, and what the current state of the art is. Both audiences should get pointers for further reading wherever the scope of this survey (defined in Section 1.2

right next) ends. But we also provide explanations of the underlying concepts and technologies that are necessary to understand the various approaches. Thus, this survey should also make a good tutorial for a researcher previously unfamiliar with semantic search.

1.2 Scope of this Survey

1.2.1 Kinds of Data

This survey focuses on semantic search on text (in natural language) or knowledge bases (consisting of structured records). The two may also be combined. For example, a natural language text may be enriched with semantic markup that identifies mentions of entities from a knowledge base. Or several knowledge bases with different schemata may be combined, like in the Semantic Web. The types of data considered in this survey are explained in detail in Section 2.1 on *Data Types and Common Datasets*.

This survey does *not* cover search on images, audio, video, and other objects that have an inherently non-textual representation. This is not to say that semantic search is not relevant for this kind of data; quite the opposite is true. For example, consider a user looking for a picture of a particular person. Almost surely, the user is not interested in the precise arrangements of pixels that are used to represent the picture. She might not even be interested in the particular angle, selection, or lighting conditions of the picture, but only in the object shown. This is very much "semantic search", but on a different kind of data. There is some overlap with search in textual data, including attempts to map non-textual to textual features and the use of text that accompanies the non-textual object (e.g., the caption of an image). But mostly, search in non-textual data is a different world that requires quite different techniques and tools.

A special case of image and audio data are scans of text documents and speech. The underlying data is also textual[4] and can be extracted using optical character recognition (OCR) and automatic speech recognition (ASR) techniques. We do not consider these techniques in this

[4]Leaving aside aspects like a particular writing style or emotions when talking.

survey. However, we acknowledge that "semantic techniques", as described in this survey, can be helpful in the text recognition process. For example, in both OCR and ASR, a semantic understanding of the possible textual interpretations can help to decide which interpretation is the most appropriate.

1.2.2 Kinds of Search

There are three types of queries prevailing in semantic search: keyword, structured, and natural language. We cover the whole spectrum in this survey; see Section 2.2 on *Search Paradigms*.

Concerning the kind of results returned, we take a narrower view: we focus on techniques and systems that are *extractive* in the sense that they return elements or excerpts from the original data. Think of the result screen from a typical web search engine. The results are nicely arranged and partly reformatted, so that we can digest them properly. But it's all excerpts and elements from the web pages and knowledge bases being searched in the background.

We only barely touch upon the analysis of query logs (queries asked) and clickthrough data (results clicked). Such data can be used to derive information on what users found relevant for a particular query. Modern web search engines leverage such information to a significant extent. This topic is out of scope for this survey, since an explicit "understanding" of the query or the data is not necessary. We refer the user to the seminal paper of Joachims [2002] and the recent survey of Silvestri [2010].

There is also a large body of research that involves the complex synthesis of new information, in particular, text. For example, in *automatic summarization*, the goal is to summarize a given (long) text document, preserving the main content and a consistent style. In *multi-document summarization*, this task is extended to multiple documents on a particular topic or question. For example, *compile a report on drug trafficking in the united states over the past decade.* Apart from collecting the various bits and pieces of text and knowledge required to answer these questions, the main challenge becomes to compile these into a compact and coherent text that is well comprehensible for humans. Such non-trivial automatic content synthesis is out of scope for this survey.

1.2.3 Further inclusion criteria

As just explained, we focus on semantic search on text and knowledge bases that retrieves elements and excerpts from the original data. But even there we cannot possibly cover all existing research in depth.

Our inclusion criteria for this survey are very practically oriented, with a focus on fundamental techniques, datasets, benchmarks, and systems. Systems were selected with a strong preference for those evaluated on one of the prevailing benchmarks or that come with a working software or demo. We provide quantitative information (on the benchmarks and the performance and effectiveness of the various systems) wherever possible.

We omit most of the history and mostly focus on the state of the art. The historical perspective is interesting and worthwhile in its own right, but the survey is already long and worthwhile without this. However, we usually mention the first system of a particular kind. Also, for each of our nine categories (explained right next, in Section 1.3), we describe systems in chronological order and make sure to clarify the improvements of the newer systems over the older ones.

1.2.4 Further Reading

The survey provides pointers for further reading at many places. Additionally, we provide here a list of well-known conferences and journals, grouped by research community, which are generally good sources for published research on the topic of this survey and beyond. In particular, the bibliography of this survey contains (many) references from each of these venues. This list is by no means complete, and there are many good papers that are right on topic but published in other venues.

Information Retrieval: SIGIR, CIKM, TREC, TAC, FNTIR.

Web and Semantic Web: WWW, ISWC, ESWC, AAAI, JWS.

Computer linguistics: ACL, EMNLP, HLT-NAACL.

Databases / Data Mining: VLDB, KDD, SIGMOD, TKDE.

1.3 Overview of this Survey

Section 1.4 provides a *Glossary* of terms that are strongly related to semantic search. For each of these, we provide a brief description together with a pointer to the relevant passages in the survey. This is useful for readers who specifically look for material on a particular problem or aspect.

Section 2 on *Classification by Data Type and Search Paradigm* describes the two main dimensions that we use for categorizing research on semantic search:

Data type: text, knowledge bases, and combined data.

Search paradigm: keyword, structured, and natural language search.

For each data type, we provide a brief characterization and a list of frequently used datasets. For each search paradigm, we provide a brief characterization and one or two examples.

Section 3 on *Basic NLP Tasks in Semantic Search* gives an overview of: part-of-speech (POS) tagging, named-entity recognition and disambiguation (NER+NED), parsing the grammatical structure of sentences, and word vectors / embeddings. These are used as basic building blocks by various (though not all) of the approaches described in our main Section 4. We give a brief tutorial on each of these tasks, as well as a succinct summary of the state of the art.

Section 4 on *Approaches and Systems for Semantic Search* is the core section of this survey. We group the many approaches and systems that exist in the literature by data type (three categories, see above) and search paradigm (three categories, see above). The resulting nine combinations are shown in Figure 1.1. In a sense, this figure is the main signpost for this survey. Note that we use *Natural Language Search* and *Question Answering* synonymously in this survey. All nine subsections share the same sub-structure:

Profile ... a short characterization of this line of research
Techniques ... what are the basic techniques used
Systems ... a concise description of milestone systems or software
Benchmarks ... existing benchmarks and the best results on them

	Keyword Search	Structured Search	Natural Lang. Search
Text	**Section 4.1** Keyword Search on Text	**Section 4.3** Structured Data Extraction from Text	**Section 4.7** Question Answering on Text
Knowledge Bases	**Section 4.4** Keyword Search on Knowledge Bases	**Section 4.2** Structured Search on Knowledge Bases	**Section 4.8** Question Answering on Knowledge Bases
Combined Data	**Section 4.5** Keyword Search on Combined Data	**Section 4.6** Semi-Struct. Search on Combined Data	**Section 4.9** Question Answering on Combined Data

Figure 1.1: Our basic classification of research on semantic search by underlying data (rows) and search paradigm (columns). The three data types are explained in Section 2.1, the three search paradigms are explained in Section 2.2. Each of the nine groups is discussed in the indicated subsection of our main Section 4.

Section 5 on *Advanced Techniques for Semantic Search* deals with: *ranking* (in semantic entity search), *indexing* (getting not only good results but getting them fast), *ontology matching and merging* (dealing with multiple knowledge bases), and *inference* (information that is not directly contained in the data but can be inferred from it). They provide a deeper understanding of the aspects that are critical for results of high quality and/or with high performance.

Section 6 on *The Future of Semantic Search* provides a very brief summary of the state of the art in semantic search, as described in the main sections of this survey, and then dares to take a look into the near and the not so near future.

The article closes with a long list of 218 references. Datasets and standards are not listed as part of the References but separately in the Appendices. In the PDF of this article, all citations in the text are clickable (leading to the respective entry in the References), and so are most of the titles in the References (leading to the respective article on the Web). In most PDF readers, *Alt+Left* brings you back to the place of the citation.

The reader may wonder about possible reading orders and which sections depend upon which. In fact, each of the six sections of this survey is relatively self-contained and readable on its own. This is true even for each of the nine subsections (one for each kind of semantic search, according to our basic classification) of the main Section 4. However, when reading such a subsection individually, it is a good idea to prepend a quick read of those subsections from Section 2 that deal with the respective data type and search paradigm: they are short and easy to read, with instructive examples. Readers looking for specific information may find the glossary, which comes right next, useful.

1.4 Glossary

This glossary provides a list of techniques or aspects that are strongly related to semantic search but non-trivial to find using our basic classification. For each item, we provide a very short description and a pointer to the relevant section(s) of the survey.

Deep learning for NLP: natural language processing using (deep) neural networks; used for the word vectors in Section 3.4; some of the systems in Section 4.8 on *Question Answering on Knowledge Bases* use deep learning or word vectors; apart from that, deep NLP is still used very little in actual systems for semantic search, but see Section 6 on *The Future of Semantic Search*.

Distant supervision: technique to derive labeled training data using heuristics in order to learn a (supervised) classifier; the basic principle and significance for semantic search is explained in Section 4.3.2 on *Systems for Relationship Extraction from Text*.

Entity resolution: identify that two different strings refer to the same entity; this is used in Section 4.3.4 on *Knowledge Base Construction* and discussed more generally in Section 5.4 on *Ontology Matching and Merging*.

Entity search/retrieval: search on text or combined data that aims at a particular entity or list of entities as opposed to a list of documents; this applies to almost all the systems in Section 4 that work with combined data or natural language queries[5]; see also Section 5.1, which is all about ranking techniques for entity search.

Knowledge base construction: constructing or enriching a knowledge base from a given text corpus; basic techniques are explained in Section 4.3.1; systems are described in Section 4.3.4.

Learning to rank for semantic search: supervised learning of good ranking functions; several applications in the context of semantic search are described in Section 5.1.

Ontology merging and matching: reconciling and aligning naming schemes and contents of different knowledge bases; this is the topic of Section 5.3.

Paraphrasing or synonyms: identifying whether two words, phrases or sentences are synonymous; systems in Section 4.8 on *Question Answering on Knowledge Bases* make use of this; three datasets that are used by systems described in this survey are: Patty [2013] (paraphrases extracted in an unsupervised fashion), Paralex [2013] (question paraphrases), and CrossWikis [2012] (Wikipedia entity anchors in multiple languages).

Question answering: synonymous with natural language search in this survey; see Section 2.2.3 for a definition; see Sections 4.7, 4.8, and 4.9 for research on question answering on each of our three data types.

Reasoning/Inference: using reasoning to infer new triples from a given knowledge base; this is the topic of Section 5.4.

Semantic parsing: finding the logical structure of a natural language query; this is described in Sections 4.8 on *Question Answering on Knowledge Bases* and used by many of the systems there.

Semantic web: a framework for explicit semantic data on the web; this kind of data is described in Section 2.1.3; the systems described

[5] A search on a knowledge base naturally returns a list of entities, too. However, the name *entity search* is usually only used when (also) text is involved and returning lists of entities is not the only option.

in Section 4.5 deal with this kind of data; it is important to note that many papers / systems that claim to be about semantic web data are actually dealing only with a single knowledge base (like DBpedia, see Table 2.2), and are hence described in the sections dealing with search on knowledge bases.

Information extraction: extracting structured information from text; this is exactly what Section 4.3 on *Structured Data Extraction from Text* is about.

XML retrieval: search in nested semi-structured data (text with tag pairs, which can be arbitrarily nested); the relevance for semantic search is discussed in Section 4.5.3 in the context of the INEX series of benchmarks.

References

Abadi, D., P. Boncz, S. Harizopoulus, S. Idreos, and S. Madden (2013). The design and implementation of modern column-oriented database systems. In: *Foundations and Trends in Databases* 5.3, pp. 197–280.

Agarwal, A., S. Chakrabarti, and S. Aggarwal (2006). Learning to rank networked entities. In: *KDD*, pp. 14–23.

Agrawal, S., S. Chaudhuri, and G. Das (2002). DBXplorer: enabling keyword search over relational databases. In: *SIGMOD*, p. 627.

Angeli, G., S. Gupta, M. Jose, C. D. Manning, C. Ré, J. Tibshirani, J. Y. Wu, S. Wu, and C. Zhang (2014). Stanford's 2014 slot filling systems. In: *TAC-KBP*.

Arasu, A. and H. Garcia-Molina (2003). Extracting structured data from web pages. In: *SIGMOD*, pp. 337–348.

Armstrong, T. G., A. Moffat, W. Webber, and J. Zobel (2009a). Has adhoc retrieval improved since 1994? In: *SIGIR*, pp. 692–693.

Armstrong, T. G., A. Moffat, W. Webber, and J. Zobel (2009b). Improvements that don't add up: ad-hoc retrieval results since 1998. In: *CIKM*, pp. 601–610.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives (2007). DBpedia: a nucleus for a web of open data. In: *ISWC/ASWC*, pp. 722–735.

Balakrishnan, S., A. Halevy, B. Harb, H. Lee, J. Madhavan, A. Rostamizadeh, W. Shen, K. Wilder, F. Wu, and C. Yu (2015). Applying WebTables in practice. In: *CIDR*.

Balmin, A., V. Hristidis, and Y. Papakonstantinou (2004). ObjectRank: authority-based keyword search in databases. In: *VLDB*, pp. 564–575.

Balog, K. and R. Neumayer (2013). A test collection for entity search in DBpedia. In: *SIGIR*, pp. 737–740.

Balog, K., P. Serdyukov, and A. P. de Vries (2010). Overview of the TREC 2010 Entity Track. In: *TREC*.

Balog, K., P. Serdyukov, and A. P. de Vries (2011). Overview of the TREC 2011 Entity Track. In: *TREC*.

Balog, K., A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld (2009). Overview of the TREC 2009 Entity Track. In: *TREC*.

Balog, K., Y. Fang, M. de Rijke, P. Serdyukov, and L. Si (2012). Expertise retrieval. In: *Foundations and Trends in Information Retrieval* 6.2-3, pp. 127–256.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007). Open information extraction from the Web. In: *IJCAI*, pp. 2670–2676.

Bast, H. and I. Weber (2006). Type less, find more: fast autocompletion search with a succinct index. In: *SIGIR*, pp. 364–371.

Bast, H., A. Chitea, F. M. Suchanek, and I. Weber (2007). ESTER: efficient search on text, entities, and relations. In: *SIGIR*, pp. 671–678.

Bast, H. and B. Buchhold (2013). An index for efficient semantic full-text search. In: *CIKM*, pp. 369–378.

Bast, H., B. Buchhold, and E. Haussmann (2015). Relevance scores for triples from type-like relations. In: *SIGIR*, pp. 243–252.

Bast, H. and E. Haussmann (2013). Open information extraction via contextual sentence decomposition. In: *ICSC*, pp. 154–159.

Bast, H. and E. Haussmann (2014). More informative open information extraction via simple inference. In: *ECIR*, pp. 585–590.

Bast, H. and E. Haussmann (2015). More accurate question answering on Freebase. In: *CIKM*, pp. 1431–1440.

Bast, H., F. Bäurle, B. Buchhold, and E. Haussmann (2012). Broccoli: semantic full-text search at your fingertips. In: *CoRR* abs/1207.2615.

Bast, H., F. Bäurle, B. Buchhold, and E. Haußmann (2014a). Easy access to the Freebase dataset. In: *WWW*, pp. 95–98.

Bast, H., F. Bäurle, B. Buchhold, and E. Haußmann (2014b). Semantic full-text search with Broccoli. In: *SIGIR*, pp. 1265–1266.

Bastings, J. and K. Sima'an (2014). All fragments count in parser evaluation. In: *LREC*, pp. 78–82.

Berant, J. and P. Liang (2014). Semantic parsing via paraphrasing. In: *ACL*, pp. 1415–1425.

Berant, J., A. Chou, R. Frostig, and P. Liang (2013a). Semantic parsing on freebase from question-answer pairs. In: *EMNLP*, pp. 1533–1544.

Berant, J., A. Chou, R. Frostig, and P. Liang (2013b). The WebQuestions Benchmark. In: Introduced by [Berant, Chou, Frostig, and Liang, 2013a].

Bhalotia, G., A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan (2002). Keyword searching and browsing in databases using BANKS. In: *ICDE*, pp. 431–440.

Bizer, C. and A. Schultz (2009). The Berlin SPARQL Benchmark. In: *IJSWIS* 5.2, pp. 1–24.

Blanco, R., P. Mika, and S. Vigna (2011). Effective and efficient entity search in RDF data. In: *ISWC*, pp. 83–97.

Blanco, R., H. Halpin, D. M. Herzig, P. Mika, J. Pound, H. S. Thompson, and D. T. Tran (2011). Entity search evaluation over structured web data. In: *SIGIR-EOS*. Vol. 2011.

Bleiholder, J. and F. Naumann (2008). Data fusion. In: *ACM Comput. Surv.* 41.1, 1:1–1:41.

Boldi, P. and S. Vigna (2005). MG4J at TREC 2005. In: *TREC*.

Bollacker, K. D., C. Evans, P. Paritosh, T. Sturge, and J. Taylor (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In: *SIGMOD*, pp. 1247–1250.

Bordes, A., S. Chopra, and J. Weston (2014). Question answering with subgraph embeddings. In: *CoRR* abs/1406.3676.

Bordes, A. and E. Gabrilovich (2015). Constructing and mining web-scale knowledge graphs: WWW 2015 tutorial. In: *WWW*, p. 1523.

Broekstra, J., A. Kampman, and F. van Harmelen (2002). Sesame: a generic architecture for storing and querying RDF and RDF schema. In: *ISWC*, pp. 54–68.

Bruijn, J. de, M. Ehrig, C. Feier, F. Martín-Recuerda, F. Scharffe, and M. Weiten (2006). Ontology mediation, merging and aligning. In: *Semantic Web Technologies*, pp. 95–113.

Bruni, E., N. Tran, and M. Baroni (2014). Multimodal Distributional Semantics. In: *JAIR* 49, pp. 1–47.

Cafarella, M., A. Halevy, D. Wang, E. Wu, and Y. Zhang (2008). WebTables: exploring the power of tables on the web. In: *PVLDB* 1.1, pp. 538–549.

Cai, Q. and A. Yates (2013). Large-scale semantic parsing via schema matching and lexicon extension. In: *ACL*, pp. 423–433.

Carlson, A., J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell (2010). Toward an architecture for never-ending language learning. In: *AAAI*, pp. 1306–1313.

Carmel, D., M.-W. Chang, E. Gabrilovich, B.-J. P. Hsu, and K. Wang (2014). ERD'14: entity recognition and disambiguation challenge. In: *SIGIR*, p. 1292.

Castano, S., A. Ferrara, S. Montanelli, and G. Varese (2011). Ontology and instance matching. In: *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, pp. 167–195.

Charniak, E. (2000). A maximum-entropy-inspired parser. In: *ANLP*, pp. 132–139.

Chen, D. and C. D. Manning (2014). A fast and accurate dependency parser using neural networks. In: *ACL*, pp. 740–750.

Cheng, G., W. Ge, and Y. Qu (2008). Falcons: searching and browsing entities on the semantic web. In: *WWW*, pp. 1101–1102.

Choi, J. D., J. R. Tetreault, and A. Stent (2015). It depends: dependency parser comparison using a web-based evaluation tool. In: *ACL*, pp. 387–396.

Cimiano, P., V. Lopez, C. Unger, E. Cabrio, A.-C. N. Ngomo, and S. Walter (2013). Multilingual question answering over linked data (QALD-3): lab overview. In: *CLEF*, pp. 321–332.

Coffman, J. and A. C. Weaver (2010). A framework for evaluating database keyword search strategies. In: *CIKM*, pp. 729–738.

Coffman, J. and A. C. Weaver (2014). An empirical performance evaluation of relational keyword search techniques. In: *TKDE* 26.1, pp. 30–42.

Cornolti, M., P. Ferragina, M. Ciaramita, H. Schütze, and S. Rüd (2014). The SMAPH system for query entity recognition and disambiguation. In: *ERD*, pp. 25–30.

Corro, L. D. and R. Gemulla (2013). ClausIE: clause-based open information extraction. In: *WWW*, pp. 355–366.

Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery (1998). Learning to extract symbolic knowledge from the world wide web. In: *AAAI*, pp. 509–516.

Cucerzan, S. (2012). The MSR system for entity linking at TAC 2012. In: *TAC*.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In: *EMNLP-CoNLL*, pp. 708–716.

Cucerzan, S. (2014). Name entities made obvious: the participation in the ERD 2014 evaluation. In: *ERD*, pp. 95–100.

Dang, H. T., D. Kelly, and J. J. Lin (2007). Overview of the TREC 2007 Question Answering Track. In: *TREC*.

Dang, H. T., J. J. Lin, and D. Kelly (2006). Overview of the TREC 2006 Question Answering Track. In: *TREC*.

Delbru, R., S. Campinas, and G. Tummarello (2012). Searching web data: An entity retrieval and high-performance indexing model. In: *J. Web Sem.* 10, pp. 33–58.

Dill, S., N. Eiron, D. Gibson, D. Gruhl, R. V. Guha, A. Jhingran, T. Kanungo, K. S. McCurley, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien (2003). A case for automated large-scale semantic annotation. In: *J. Web Sem.* 1.1, pp. 115–132.

Ding, L., T. W. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs (2004). Swoogle: a search and metadata engine for the semantic web. In: *CIKM*, pp. 652–659.

Doan, A. and A. Y. Halevy (2005). Semantic integration research in the database community: a brief survey. In: *AI Magazine*, pp. 83–94.

Dong, X., E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang (2014). Knowledge Vault: a web-scale approach to probabilistic knowledge fusion. In: *KDD*, pp. 601–610.

Elbassuoni, S., M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum (2009). Language-model-based ranking for queries on RDF-graphs. In: *CIKM*, pp. 977–986.

Elliott, B., E. Cheng, C. Thomas-Ogbuji, and Z. M. Özsoyoglu (2009). A complete translation from SPARQL into efficient SQL. In: *IDEAS*, pp. 31–42.

Elmagarmid, A. K., P. G. Ipeirotis, and V. S. Verykios (2007). Duplicate record detection: a survey. In: *TKDE*, pp. 1–16.

Etzioni, O., A. Fader, J. Christensen, S. Soderland, and Mausam (2011). Open information extraction: the second generation. In: *IJCAI*, pp. 3–10.

Euzenat, J., A. Ferrara, C. Meilicke, J. Pane, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, and C. dos Santos (2010). Results of the ontology alignment evaluation initiative 2010. In: *OM*, pp. 85–117.

Euzenat, J., C. Meilicke, H. Stuckenschmidt, P. Shvaiko, and C. dos Santos (2011a). Ontology alignment evaluation initiative: six years of experience. In: *J. Data Semantics* 15, pp. 158–192.

Euzenat, J., A. Ferrara, W. R. van Hage, L. Hollink, C. Meilicke, A. Nikolov, D. Ritze, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, and C. dos Santos (2011b). Results of the ontology alignment evaluation initiative 2011. In: *OM*, pp. 158–192.

Fader, A., S. Soderland, and O. Etzioni (2011). Identifying relations for open information extraction. In: *EMNLP*, pp. 1535–1545.

Fader, A., L. S. Zettlemoyer, and O. Etzioni (2013). Paraphrase-driven learning for open question answering. In: *ACL*, pp. 1608–1618.

Fang, Y., L. Si, Z. Yu, Y. Xian, and Y. Xu (2009). Entity retrieval with hierarchical relevance model, exploiting the structure of tables and learning homepage classifiers. In: *TREC*.

Ferrucci, D. A., E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer, and C. A. Welty (2010). Building Watson: An Overview of the DeepQA Project. In: *AI Magazine* 31.3, pp. 59–79.

Ferrucci, D. A., A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller (2013). Watson: Beyond Jeopardy! In: *Artif. Intell.* 199, pp. 93–105.

Finkel, J. R., T. Grenager, and C. D. Manning (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In: *ACL*, pp. 363–370.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin (2002). Placing search in context: the concept revisited. In: *TOIS* 20.1, pp. 116–131.

Frank, J., S. Bauer, M. Kleiman-Weiner, D. Roberts, N. Tripuaneni, C. Zhang, C. Ré, E. Voorhees, and I. Soboroff (2013). Stream filtering for entity profile updates for TREC 2013. In: *TREC-KBA*.

Frank, J., M. Kleiman-Weiner, D. A. Roberts, E. Voorhees, and I. Soboroff (2014). Evaluating stream filtering for entity profile updates in TREC 2012, 2013, and 2014. In: *TREC-KBA*.

Frank, J. R., M. Kleiman-Weiner, D. A. Roberts, F. Niu, C. Zhang, C. Ré, and I. Soboroff (2012). Building an entity-centric stream filtering test collection for TREC 2012. In: *TREC-KBA*.

Franz, T., A. Schultz, S. Sizov, and S. Staab (2009). TripleRank: ranking semantic web data by tensor decomposition. In: *ISWC*, pp. 213–228.

Fundel, K., R. Küffner, and R. Zimmer (2007). RelEx - relation extraction using dependency parse trees. In: *Bioinformatics* 23.3, pp. 365–371.

Gabrilovich, E. and S. Markovitch (2007). Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In: *IJCAI*. Vol. 7, pp. 1606–1611.

Gaifman, H. (1965). Dependency systems and phrase-structure systems. In: *Information and Control* 8.3, pp. 304–337.

Gövert, N., N. Fuhr, M. Lalmas, and G. Kazai (2006). Evaluating the effectiveness of content-oriented XML retrieval methods. In: *Information Retrieval* 9.6, pp. 699–722.

Grau, B. C., I. Horrocks, B. Motik, B. Parsia, P. F. Patel-Schneider, and U. Sattler (2008). OWL 2: the next step for OWL. In: *J. Web Sem.* 6.4, pp. 309–322.

Grau, B. C., Z. Dragisic, K. Eckert, J. Euzenat, A. Ferrara, R. Granada, V. Ivanova, E. Jiménez-Ruiz, A. O. Kempf, P. Lambrix, A. Nikolov, H. Paulheim, D. Ritze, F. Scharffe, P. Shvaiko, C. T. dos Santos, and O. Zamazal (2013). Results of the ontology alignment evaluation initiative 2013. In: *OM*, pp. 61–100.

Guha, R. V., R. McCool, and E. Miller (2003). Semantic search. In: *WWW*, pp. 700–709.

Guha, R., D. Brickley, and S. MacBeth (2015). Schema.org: evolution of structured data on the web. In: *ACM Queue* 13.9, p. 10.

Guo, Y., Z. Pan, and J. Heflin (2005). LUBM: a benchmark for OWL knowledge base systems. In: *J. Web Sem.* 3, pp. 158–182.

Halpin, H., D. Herzig, P. Mika, R. Blanco, J. Pound, H. Thompson, and D. T. Tran (2010). Evaluating ad-hoc object retrieval. In: *IWEST*.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In: *COLING*, pp. 539–545.

Herzig, D. M., P. Mika, R. Blanco, and T. Tran (2013). Federated entity search using on-the-fly consolidation. In: *ISWC*, pp. 167–183.

Hill, F., R. Reichart, and A. Korhonen (2015). SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. In: *Computational Linguistics* 41.4, pp. 665–695.

Hoffart, J., F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum (2011). YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In: *WWW*, pp. 229–232.

Hoffart, J., F. M. Suchanek, K. Berberich, and G. Weikum (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In: *Artif. Intell.* 194, pp. 28–61.

Hoffmann, R., C. Zhang, X. Ling, L. S. Zettlemoyer, and D. S. Weld (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In: *ACL*, pp. 541–550.

Hovy, E. H., M. P. Marcus, M. Palmer, L. A. Ramshaw, and R. M. Weischedel (2006). OntoNotes: the 90% solution. In: *HLT-NAACL*, pp. 57–60.

Hristidis, V. and Y. Papakonstantinou (2002). DISCOVER: keyword search in relational databases. In: *VLDB*, pp. 670–681.

Hua, W., Z. Wang, H. Wang, K. Zheng, and X. Zhou (2015). Short text understanding through lexical-semantic analysis. In: *ICDE*, pp. 495–506.

Ji, H., R. Grishman, and H. T. Dang (2011). Overview of the TAC 2011 Knowledge Base Population Track. In: *TAC-KBP*.

Ji, H., J. Nothman, and B. Hachey (2014). Overview of TAC-KBP 2014 entity discovery and linking tasks. In: *TAC-KBP*.

Ji, H., R. Grishman, H. T. Dang, K. Griffit, and J. Ellisa (2010). Overview of the TAC 2010 Knowledge Base Population Track. In: *TAC-KBP*.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In: *KDD*, pp. 133–142.

Joshi, M., U. Sawant, and S. Chakrabarti (2014). Knowledge graph and corpus driven segmentation and answer inference for telegraphic entity-seeking queries. In: *EMNLP*, pp. 1104–1114.

Kaptein, R. and J. Kamps (2013). Exploiting the category structure of Wikipedia for entity ranking. In: *Artif. Intell.* 194, pp. 111–129.

Katz, B. (1997). Annotating the world wide web using natural language. In: *RIAO*, pp. 136–159.

Katz, B., G. C. Borchardt, and S. Felshin (2006). Natural language annotations for question answering. In: *FLAIRS*, pp. 303–306.

Klein, D. and C. D. Manning (2002). Fast exact inference with a factored model for natural language parsing. In: *NIPS*, pp. 3–10.

Kolomiyets, O. and M. Moens (2011). A survey on question answering technology from an information retrieval perspective. In: *Inf. Sci.* 181.24, pp. 5412–5434.

Köpcke, H. and E. Rahm (2010). Frameworks for entity matching: a comparison. In: *DKE*, pp. 197–210.

Le, Q. V. and T. Mikolov (2014). Distributed representations of sentences and documents. In: *ICML*, pp. 1188–1196.

Lee, D. D. and H. S. Seung (2000). Algorithms for non-negative matrix factorization. In: *NIPS*, pp. 556–562.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. In: *Semantic Web* 6.2, pp. 167–195.

Lei, Y., V. S. Uren, and E. Motta (2006). SemSearch: a search engine for the semantic web. In: *EKAW*, pp. 238–245.

Levy, O. and Y. Goldberg (2014). Neural word embedding as implicit matrix factorization. In: *NIPS*, pp. 2177–2185.

Levy, O., Y. Goldberg, and I. Dagan (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. In: *TACL* 3, pp. 211–225.

Li, G., S. Ji, C. Li, and J. Feng (2009). Efficient type-ahead search on relational data: a TASTIER approach. In: *SIGMOD*, pp. 695–706.

Li, H. and J. Xu (2014). Semantic matching in search. In: *Foundations and Trends in Information Retrieval* 7.5, pp. 343–469.

Limaye, G., S. Sarawagi, and S. Chakrabarti (2010). Annotating and Searching Web Tables Using Entities, Types and Relationships. In: *PVLDB* 3.1, pp. 1338–1347.

Liu, T. (2009). Learning to rank for information retrieval. In: *Foundations and Trends in Information Retrieval* 3.3, pp. 225–331.

Lopez, V., V. S. Uren, M. Sabou, and E. Motta (2011a). Is question answering fit for the Semantic Web?: A survey. In: *Semantic Web* 2.2, pp. 125–155.

Lopez, V., C. Unger, P. Cimiano, and E. Motta (2011b). Proceedings of the 1st workshop on question answering over linked data (QALD-1). In: *ESWC*.

Lopez, V., C. Unger, P. Cimiano, and E. Motta (2012). Interacting with linked data. In: *ESWC-ILD*.

Lopez, V., C. Unger, P. Cimiano, and E. Motta (2013). Evaluating question answering over linked data. In: *J. Web Sem.* 21, pp. 3–13.

Lund, K. and C. Burgess (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. In: *Behavior research methods, instruments, & computers* 28.2, pp. 203–208.

Luong, T., R. Socher, and C. D. Manning (2013). Better word representations with recursive neural networks for morphology. In: *CoNLL*, pp. 104–113.

Ma, L., Y. Yang, Z. Qiu, G. T. Xie, Y. Pan, and S. Liu (2006). Towards a complete OWL ontology benchmark. In: *ESWC*, pp. 125–139.

Macdonald, C. and I. Ounis (2006). The TREC Blogs06 collection: Creating and analysing a blog test collection. In: *Department of Computer Science, University of Glasgow Tech Report TR-2006-224* 1, pp. 3–1.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: *CICLING*, pp. 171–189.

Manning, C. D., M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In: *ACL*, pp. 55–60.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: the Penn Treebank. In: *Computational Linguistics* 19.2, pp. 313–330.

Mass, Y. and Y. Sagiv (2012). Language models for keyword search over data graphs. In: *WSDM*, pp. 363–372.

Mausam, M. Schmitz, S. Soderland, R. Bart, and O. Etzioni (2012). Open language learning for information extraction. In: *EMNLP-CoNLL*, pp. 523–534.

Mayfield, J., J. Artiles, and H. T. Dang (2012). Overview of the TAC 2012 Knowledge Base Population Track. In: *TAC-KBP*.

Mayfield, J. and R. Grishman (2015). TAC 2015 Cold Start KBP Track. In: *TAC-KBP*.

McClosky, D., E. Charniak, and M. Johnson (2006). Effective self-training for parsing. In: *HLT-NAACL*.

Meusel, R., P. Petrovski, and C. Bizer (2014). The WebDataCommons Microdata, RDFa and Microformat dataset series. In: *ISWC*, pp. 277–292.

Mikolov, T., W. Yih, and G. Zweig (2013). Linguistic regularities in continuous space word representations. In: *NAACL*, pp. 746–751.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013a). Distributed representations of words and phrases and their compositionality. In: *NIPS*, pp. 3111–3119.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013b). Efficient estimation of word representations in vector space. In: *CoRR* abs/1301.3781.

Miller, G. A. (1992). WordNet: A Lexical Database for English. In: *Commun. ACM* 38, pp. 39–41.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky (2009). Distant supervision for relation extraction without labeled data. In: *ACL/IJCNLP*, pp. 1003–1011.

Mitchell, T. M., W. W. Cohen, E. R. H. Jr., P. P. Talukdar, J. Betteridge, A. Carlson, B. D. Mishra, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. A. Platanios, A. Ritter, M. Samadi, B. Settles, R. C. Wang, D. T. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling (2015). Never-ending learning. In: *AAAI*, pp. 2302–2310.

Mitkov, R. (2014). *Anaphora resolution*. Routledge.

Moldovan, D. I., C. Clark, and M. Bowden (2007). Lymba's PowerAnswer 4 in TREC 2007. In: *TREC*.

Monahan, S., D. Carpenter, M. Gorelkin, K. Crosby, and M. Brunson (2014). Populating a knowledge base with entities and events. In: *TAC*.

Morsey, M., J. Lehmann, S. Auer, and A.-C. N. Ngomo (2011). DBpedia SPARQL benchmark - performance assessment with real queries on real data. In: *ISWC*, pp. 454–469.

Nakashole, N., G. Weikum, and F. M. Suchanek (2012). PATTY: A taxonomy of relational patterns with semantic types. In: *EMNLP*, pp. 1135–1145.

Neumann, T. and G. Weikum (2009). Scalable join processing on very large RDF graphs. In: *SIGMOD*, pp. 627–640.

Neumann, T. and G. Weikum (2010). The RDF-3X engine for scalable management of RDF data. In: *VLDB J.* 19.1, pp. 91–113.

Neumayer, R., K. Balog, and K. Nørvåg (2012). On the modeling of entities for ad-hoc entity search in the web of data. In: *ECIR*, pp. 133–145.

Ng, V. (2010). Supervised noun phrase coreference research: the first fifteen years. In: *ACL*, pp. 1396–1411.

Nivre, J., J. Hall, S. Kübler, R. T. McDonald, J. Nilsson, S. Riedel, and D. Yuret (2007). The CoNLL 2007 Shared Task on dependency parsing. In: *EMNLP-CoNLL*, pp. 915–932.

Noy, N. F. and M. A. Musen (2000). PROMPT: algorithm and tool for automated ontology merging and alignment. In: *AAAI*, pp. 450–455.

Oren, E., R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello (2008). Sindice.com: a document-oriented lookup index for open linked data. In: *IJMSO* 3.1, pp. 37–52.

Orr, D., A. Subramanya, E. Gabrilovich, and M. Ringgaard (2013). 11 billion clues in 800 million documents: a web research corpus annotated with freebase concepts. In: *Google Research Blog*.

Park, S., S. Kwon, B. Kim, and G. G. Lee (2015). ISOFT at QALD-5: hybrid question answering system over linked data and text data. In: *CLEF*.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: global vectors for word representation. In: *EMNLP*, pp. 1532–1543.

Petrov, S. and R. McDonald (2012). Overview of the 2012 shared task on parsing the web. In: *SANCL*. Vol. 59.

Pickover, C. A., ed. (2012). *This is Watson* 56.3–4: *IBM Journal of Research and Development*.

Popov, B., A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov (2004). KIM - a semantic platform for information extraction and retrieval. In: *Natural Language Engineering* 10.3-4, pp. 375–392.

Pound, J., P. Mika, and H. Zaragoza (2010). Ad-hoc object retrieval in the web of data. In: *WWW*, pp. 771–780.

Pound, J., A. K. Hudek, I. F. Ilyas, and G. E. Weddell (2012). Interpreting keyword queries over web knowledge bases. In: *CIKM*, pp. 305–314.

Prager, J. M. (2006). Open-domain question-answering. In: *Foundations and Trends in Information Retrieval* 1.2, pp. 91–231.

Qi, Y., Y. Xu, D. Zhang, and W. Xu (2014). BUPT_PRIS at TREC 2014 knowledge base acceleration track. In: *TREC*.

Radinsky, K., E. Agichtein, E. Gabrilovich, and S. Markovitch (2011). A word at a time: computing word relatedness using temporal semantic analysis. In: *WWW*, pp. 337–346.

Reddy, S., M. Lapata, and M. Steedman (2014). Large-scale Semantic Parsing without Question-Answer Pairs. In: *TACL* 2, pp. 377–392.

Riedel, S., L. Yao, and A. McCallum (2010). Modeling relations and their mentions without labeled text. In: *ECML PKDD*, pp. 148–163.

Sarawagi, S. (2008). Information Extraction. In: *Foundations and Trends in Databases* 1.3, pp. 261–377.

Schmidt, M., M. Meier, and G. Lausen (2010). Foundations of SPARQL query optimization. In: *ICDT*, pp. 4–33.

Schuhmacher, M., L. Dietz, and S. P. Ponzetto (2015). Ranking entities for web queries through text and knowledge. In: *CIKM*, pp. 1461–1470.

Shvaiko, P. and J. Euzenat (2013). Ontology matching: state of the art and future challenges. In: *TKDE* 25.1, pp. 158–176.

Silvestri, F. (2010). Mining query logs: turning search usage data into knowledge. In: *Foundations and Trends in Information Retrieval* 4.1-2, pp. 1–174.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz (2007). Pellet: A practical OWL-DL reasoner. In: *J. Web Sem.* 5.2, pp. 51–53.

Socher, R., J. Bauer, C. D. Manning, and A. Y. Ng (2013). Parsing with compositional vector grammars. In: *ACL (1)*, pp. 455–465.

Spitkovsky, V. I. and A. X. Chang (2012). A cross-lingual dictionary for english wikipedia concepts. In: *LREC*, pp. 3168–3175.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). YAGO: a core of semantic knowledge. In: *WWW*, pp. 697–706.

Surdeanu, M. (2013). Overview of the TAC 2013 Knowledge Base Population evaluation: english slot filling and temporal slot filling. In: *TAC-KBP*.

Surdeanu, M. and H. Ji (2014). Overview of the english slot filling track at the TAC 2014 Knowledge Base Population evaluation. In: *TAC-KBP*.

Surdeanu, M., J. Tibshirani, R. Nallapati, and C. D. Manning (2012). Multi-instance multi-label learning for relation extraction. In: *EMNLP-CoNLL*, pp. 455–465.

Tablan, V., K. Bontcheva, I. Roberts, and H. Cunningham (2015). Mímir: An open-source semantic search framework for interactive information seeking and discovery. In: *J. Web Sem.* 30, pp. 52–68.

Tonon, A., G. Demartini, and P. Cudré-Mauroux (2012). Combining inverted indices and structured search for ad-hoc object retrieval. In: *SIGIR*, pp. 125–134.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In: *HLT-NAACL*, pp. 173–180.

Tran, T., H. Wang, and P. Haase (2009). Hermes: data web search on a pay-as-you-go integration infrastructure. In: *J. Web Sem.* 7.3, pp. 189–203.

Tran, T., P. Cimiano, S. Rudolph, and R. Studer (2007). Ontology-based interpretation of keywords for semantic search. In: *ISWC/ASWC*, pp. 523–536.

Trotman, A., C. L. A. Clarke, I. Ounis, S. Culpepper, M. Cartright, and S. Geva (2012). Open source information petrieval: a report on the SIGIR 2012 workshop. In: *SIGIR Forum* 46.2, pp. 95–101.

Unbehauen, J., C. Stadler, and S. Auer (2013). Optimizing SPARQL-to-SQL rewriting. In: *IIWAS*, p. 324.

Unger, C., L. Bühmann, J. Lehmann, A. N. Ngomo, D. Gerber, and P. Cimiano (2012). Template-based question answering over RDF data. In: *WWW*, pp. 639–648.

Unger, C., C. Forascu, V. Lopez, A. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter (2014). Question answering over linked data (QALD-4). In: *CLEF*, pp. 1172–1180.

Unger, C., C. Forascu, V. Lopez, A. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter (2015). Question answering over linked data (QALD-5). In: *CLEF*.

Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. In: *TREC*.

Voorhees, E. M. (2000). Overview of the TREC-9 Question Answering Track. In: *TREC*.

Voorhees, E. M. (2001). Overview of the TREC 2001 Question Answering Track. In: *TREC*.

Voorhees, E. M. (2002). Overview of the TREC 2002 Question Answering Track. In: *TREC*.

Voorhees, E. M. (2003). Overview of the TREC 2003 Question Answering Track. In: *TREC*.

Voorhees, E. M. (2004). Overview of the TREC 2004 Question Answering Track. In: *TREC*.

Voorhees, E. M. and H. T. Dang (2005). Overview of the TREC 2005 Question Answering Track. In: *TREC*.

Voorhees, E. M. and D. K. Harman (2005). *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge.

Wang, H., Q. Liu, T. Penin, L. Fu, L. Zhang, T. Tran, Y. Yu, and Y. Pan (2009). Semplore: A scalable IR approach to search the Web of Data. In: *J. Web Sem.* 7.3, pp. 177–188.

Wang, Q., J. Kamps, G. R. Camps, M. Marx, A. Schuth, M. Theobald, S. Gurajada, and A. Mishra (2012). Overview of the INEX 2012 Linked Data Track. In: *CLEF*.

Wu, S., C. Zhang, F. Wang, and C. Ré (2015). Incremental Knowledge Base Construction Using DeepDive. In: *PVLDB* 8.11, pp. 1310–1321.

Xu, K., Y. Feng, and D. Zhao (2014). Answering natural language questions via phrasal semantic parsing. In: *CLEF*, pp. 1260–1274.

Yahya, M., K. Berberich, S. Elbassuoni, M. Ramanath, V. Tresp, and G. Weikum (2012). Natural language questions for the web of data. In: *EMNLP-CoNLL 2012*, pp. 379–390.

Yates, A., M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, and S. Soderland (2007). TextRunner: open information extraction on the web. In: *HLT-NAACL*, pp. 25–26.

Yih, W., M. Chang, X. He, and J. Gao (2015). Semantic parsing via staged query graph generation: question answering with knowledge base. In: *ACL*, pp. 1321–1331.

Yu, J. X., L. Qin, and L. Chang (2010). Keyword search in relational databases: a survey. In: *IEEE Data Eng. Bull.* 33.1, pp. 67–78.

Zaragoza, H., N. Craswell, M. J. Taylor, S. Saria, and S. E. Robertson (2004). Microsoft cambridge at TREC 13: web and hard tracks. In: *TREC*.

Zelenko, D., C. Aone, and A. Richardella (2003). Kernel methods for relation extraction. In: *Journal of Machine Learning Research* 3, pp. 1083–1106.

Zenz, G., X. Zhou, E. Minack, W. Siberski, and W. Nejdl (2009). From keywords to semantic queries - Incremental query construction on the semantic web. In: *J. Web Sem.* 7.3, pp. 166–176.

Zhang, C. (2015). DeepDive: A Data Management System for Automatic Knowledge Base Construction. PhD thesis. University of Wisconsin-Madison.

Zhiltsov, N., A. Kotov, and F. Nikolaev (2015). Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In: *SIGIR*, pp. 253–262.

Zhou, G., J. Su, J. Zhang, and M. Zhang (2005). Exploring various knowledge in relation extraction. In: *ACL*, pp. 427–434.

Zhou, Q., C. Wang, M. Xiong, H. Wang, and Y. Yu (2007). SPARK: adapting keyword query to semantic search. In: *ISWC/ASWC*, pp. 694–707.