NTNU – Trondheim
Norwegian University of
Science and Technology

Université Claude Bernard Lyon 1

LIRIS

An Integrated Approach for Large-scale Relation Extraction from the Web

Naimdjon Takhirov, Fabien Duchateau,
Trond Aalberg and Ingeborg Sølvberg

APWeb'2013 Sydney, Australia
April 4th, 2013

Knowledge extraction

- creation of knowledge from structured and unstructured text
- machine processable representation
- similar to IE but goes further (backed by a schema)
- many projects towards transforming databases and other structured and unstructured text into an RDF/OWL representation

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Deutsch
Français
Italiano
Suomi
Svenska

Bored of the Rings

From Wikipedia, the free encyclopedia

This article is about the 1969 parody novel of Lord of the Rings. For the computer game, see Bored of the Rings (computer game). For The Sarah Silverman Program episode, see List of The Sarah Silverman Program episodes. For the Hughleys episode, see List of The Hughleys episodes.

Bored of the Rings is the title of a paperback parody of J. R. R. Tolkien's *The Lord of the Rings*. This short novel was written by Henry N. Beard and Douglas C. Kenney, who later founded *National Lampoon*. It was published in 1969 by Signet for the *Harvard Lampoon*.

Contents [hide]

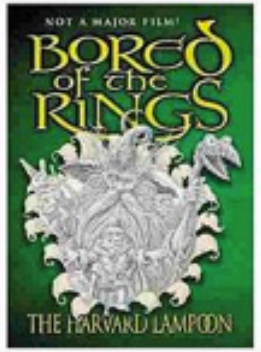
Overview [edit]

The parody generally follows the outline of *The Lord of the Rings*, including the preface, the prologue, poetry, and songs, while making light of what Tolkien made serious (e.g., "He would have finished him off then and there, but pity stayed his hand. *It's a pity I've run out of bullets*, he thought, as he went back up the tunnel..."). Names and words in the various languages are parodied with brand names which mimic their sounds (for example, *Moxie* and *Pepsi* replace *Merry* and *Pippin*). There are many topical references, including once-popular brand names. It has the distinction for a parody of having been continuously in print since it was first published.

Aside from the text itself, the book includes five elements that parody common features of mass-market books:

- A laudatory back cover review, written at Harvard, possibly by the authors themselves.
- Inside cover reviews which are entirely contrived, concluding with a quote by someone affiliated with the publication *Our Loosely Enforced Libel Laws*.
- A list of other books in the "series", none of which exist.
- A double page map which has almost nothing to do with the events in the text.
- The first text a browsing reader is liable to see purports to be a salacious sample from the book, but the episode never happens in the main text, nor does

Bored of the Rings



Front cover of the 2001 edition

Author(s)	Henry N. Beard, Douglas C. Kenney
Illustrator	William S. Donnell (map)
Cover artist	Michael K. Frith (1969 ed.) Douglas Carrel (2001 ed.)
Country	United States of America
Genre(s)	Fantasy satire
Publication date	1969
ISBN	978-0-575-07362-3

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Deutsch
Français
Italiano
Suomi
Svenska

Bored of the Rings

From Wikipedia, the free encyclopedia

This article is about the 1969 parody novel of Lord of the Rings. For the computer game, see Bored of the Rings (computer game). For The Sarah Silverman Program episode, see List of The Sarah Silverman Program episodes. For the Hughleys episode, see List of The Hughleys episodes.

Bored of the Rings is the title of a paperback parody of J. R. R. Tolkien's *The Lord of the Rings*. This short novel was written by Henry N. Beard and Douglas C. Kenney, who later founded *National Lampoon*. It was published in 1969 by Signet for the *Harvard Lampoon*.

Contents [hide]

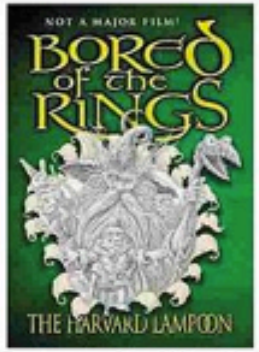
Overview [edit]

The parody generally follows the outline of *The Lord of the Rings*, including the preface, the prologue, poetry, and songs, while making light of what Tolkien made serious (e.g., "He would have finished him off then and there, but pity stayed his hand. *It's a pity I've run out of bullets*, he thought, as he went back up the tunnel..."). Names and words in the various languages are parodied with brand names which mimic their sounds (for example, *Moxie* and *Pepsi* replace *Merry* and *Pippin*). There are many topical references, including once-popular brand names. It has the distinction for a parody of having been continuously in print since it was first published.

Aside from the text itself, the book includes five elements that parody common features of mass-market books:

- A laudatory back cover review, written at Harvard, possibly by the authors themselves.
- Inside cover reviews which are entirely contrived, concluding with a quote by someone affiliated with the publication *Our Loosely Enforced Libel Laws*.
- A list of other books in the "series", none of which exist.
- A double page map which has almost nothing to do with the events in the text.
- The first text a browsing reader is liable to see purports to be a salacious sample from the book, but the episode never happens in the main text, nor does

Bored of the Rings



Front cover of the 2001 edition

Author(s)	Henry N. Beard, Douglas C. Kenney
Illustrator	William S. Donnell (map)
Cover artist	Michael K. Frith (1969 ed.) Douglas Carrel (2001 ed.)
Country	United States of America
Genre(s)	Fantasy satire
Publication date	1969
ISBN	978-0-575-07362-3

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Deutsch
Français
Italiano
Suomi
Svenska

Bored of the Rings

From Wikipedia, the free encyclopedia

This article is about the 1969 parody novel of Lord of the Rings. For the computer game, see Bored of the Rings (computer game). For The Sarah Silverman Program episode, see List of The Sarah Silverman Program episodes. For the Hughleys episode, see List of The Hughleys episodes.

Bored of the Rings is the title of a paperback parody of J. R. R. Tolkien's *The Lord of the Rings*. This short novel was written by Henry N. Beard and Douglas C. Kenney, who later founded *National Lampoon*. It was published in 1969 by Signet for the *Harvard Lampoon*.

Contents [hide]

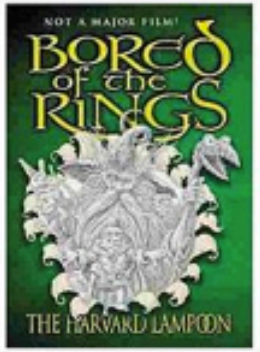
Overview [edit]

The parody generally follows the outline of *The Lord of the Rings*, including the preface, the prologue, poetry, and songs, while making light of what Tolkien made serious (e.g., "He would have finished him off then and there, but pity stayed his hand. *It's a pity I've run out of bullets*, he thought, as he went back up the tunnel..."). Names and words in the various languages are parodied with brand names which mimic their sounds (for example, *Moxie* and *Pepsi* replace *Merry* and *Pippin*). There are many topical references, including once-popular brand names. It has the distinction for a parody of having been continuously in print since it was first published.

Aside from the text itself, the book includes five elements that parody common features of mass-market books:

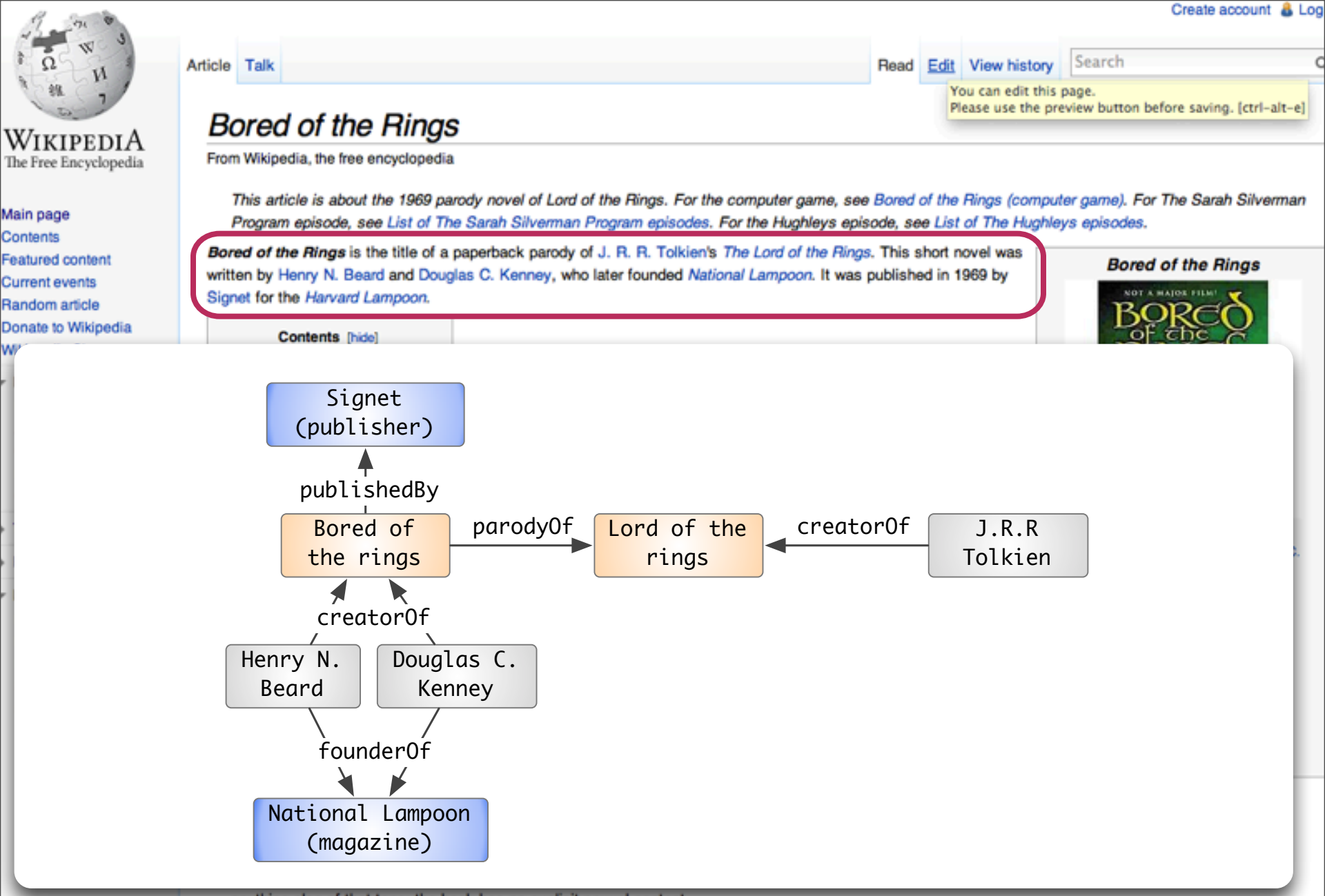
- A laudatory back cover review, written at Harvard, possibly by the authors themselves.
- Inside cover reviews which are entirely contrived, concluding with a quote by someone affiliated with the publication *Our Loosely Enforced Libel Laws*.
- A list of other books in the "series", none of which exist.
- A double page map which has almost nothing to do with the events in the text.
- The first text a browsing reader is liable to see purports to be a salacious sample from the book, but the episode never happens in the main text, nor does

Bored of the Rings



Front cover of the 2001 edition

Author(s)	Henry N. Beard, Douglas C. Kenney
Illustrator	William S. Donnell (map)
Cover artist	Michael K. Frith (1969 ed.) Douglas Carrel (2001 ed.)
Country	United States of America
Genre(s)	Fantasy satire
Publication date	1969
ISBN	978-0-575-07362-3


Create account
Log
Article
Talk
Read
Edit
View history
Search
You can edit this page.
Please use the preview button before saving. [ctrl-alt-e]
WIKIPEDIA
The Free Encyclopedia
Bored of the Rings
From Wikipedia, the free encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
This article is about the 1969 parody novel of Lord of the Rings. For the computer game, see Bored of the Rings (computer game). For The Sarah Silverman Program episode, see List of The Sarah Silverman Program episodes. For the Hughleys episode, see List of The Hughleys episodes.
Bored of the Rings is the title of a paperback parody of J. R. R. Tolkien's The Lord of the Rings. This short novel was written by Henry N. Beard and Douglas C. Kenney, who later founded National Lampoon. It was published in 1969 by Signet for the Harvard Lampoon.
Contents [hide]
Bored of the Rings
Signet (publisher)
publishedBy
Bored of the rings
parodyOf
Lord of the rings
creatorOf
J.R.R Tolkien
creatorOf
Henry N. Beard
Douglas C. Kenney
founderOf
National Lampoon (magazine)

Background (2)

- proper semantic integration of this data enables advanced semantic services (e.g. semantic and exploratory search, QA, entity matching and disambiguation, etc)
- projects: Snowball, Dipre, Espresso, NELL, ReVerb, Sofie/ Prospera, KnowItAll, Probase, etc
- also commercial interest: Google Knowledge Graph, Bing Snapshot, trueknowledge.com, etc
- issues: not typed entities/relations, multiple relations, temporal aspect, tradeoff recall/precision, runtime performance

Agenda

- context and overview
- approach
  - pattern generation
  - relationship and example generation
  - scalability
- experimental evaluation
  - relationship discovery
  - performance
- conclusion

Context

- existing, domain specific data models (e.g. libraries) need an "upgrade"
  - data created several decades ago (legacy data)
  - large investments (on the infrastructure and manpower)
- new semantic data models require a complete conversion
- recent developments of Linked Open Data(LOD) and the interest in semantic data models
- ad-hoc conversion to semantic data models (RDF, OWL etc) is difficult
  - identification of entities
  - ambiguity

Context (2)

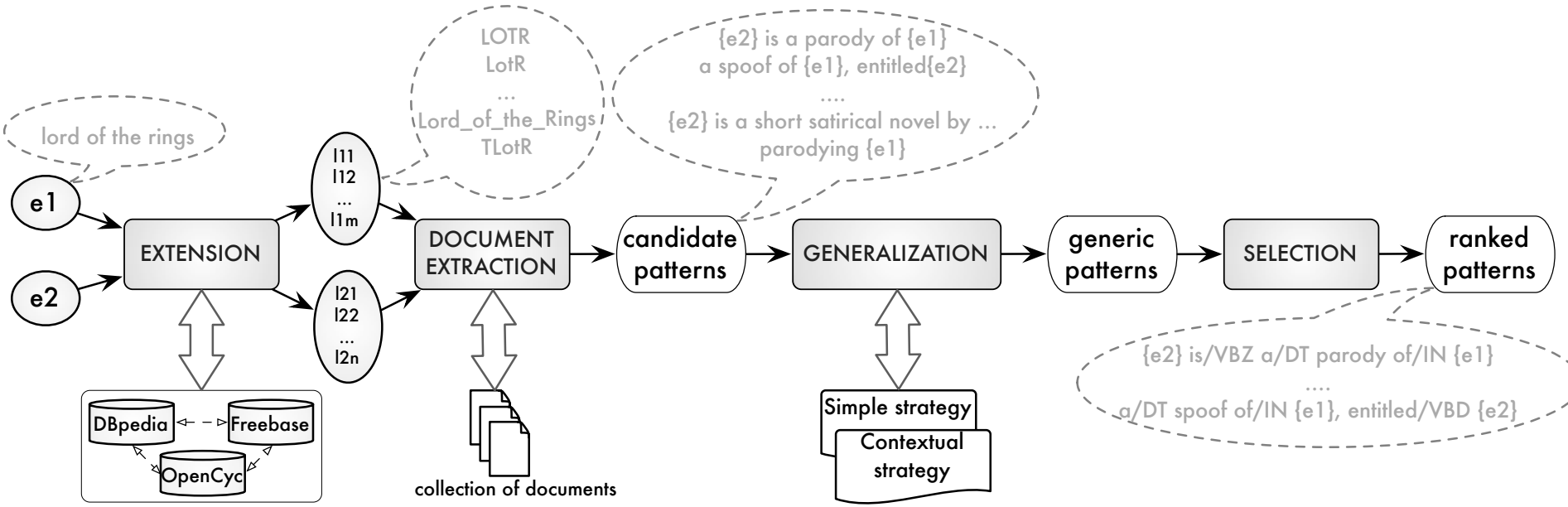
- why knowledge extraction from the Web?
  - huge source of information
    - *"Every 2 Days We Create As Much Information As We Did Up To 2003"*, E. Schmidt 2010
  - the place we discuss and share knowledge about our cultural heritage (news, wikis, blogs, personal catalogs etc.)

- SPIDER - **S**emantic and **P**rovenance-based **I**ntegration for **D**etecting and **E**xtracting **R**elations
  - extracting semantic information from the documents
  - reasonable recall/precision wrt state-of-the-art

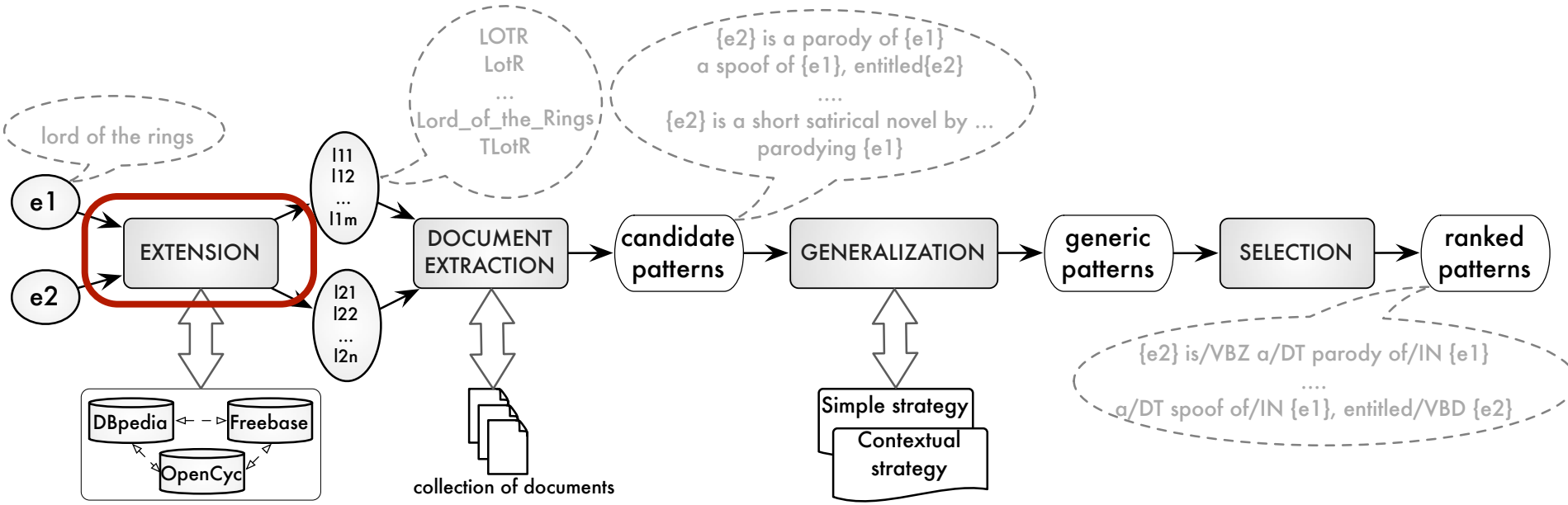
Overview

- two-step approach:
  - pattern generation
  - relationship example generation
- both patterns and examples are stored in a knowledge base

Overview of pattern generation



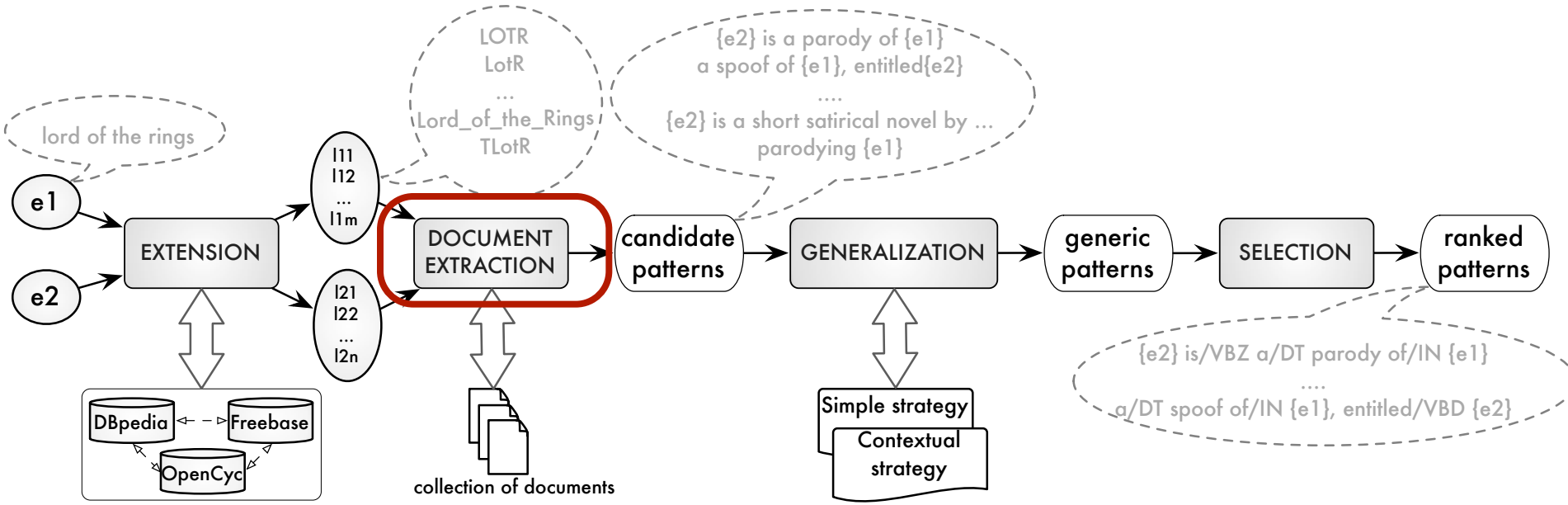
Extension


lord of the rings
e1
e2
EXTENSION
DBpedia
Freebase
OpenCyc
I11
I12
...
I1m
I21
I22
...
I2n
LOTR
LotR
...
Lord_of_the_Rings
TLotR
{e2} is a parody of {e1}
a spoof of {e1}, entitled{e2}
....
{e2} is a short satirical novel by ...
parodying {e1}
DOCUMENT EXTRACTION
collection of documents
candidate patterns
GENERALIZATION
Simple strategy
Contextual strategy
generic patterns
SELECTION
ranked patterns
{e2} is/VBZ a/DT parody of/IN {e1}
....
a/DT spoof of/IN {e1}, entitled/VBD {e2}

Extending entities

- variety of spelling forms for entities (e.g. "Lord of the Rings", "The Lord of the Rings", "LOTR" etc)
- use all alternative labels during extraction (avoid missing potentially interesting relationships)
- idea is to exploit knowledge bases (DBpedia, Freebase)
- context-driven, based on co-occurrence
- discover alternative labels in knowledge bases (e.g. dbpedia:wikiPageRedirects, freebase:common.topic.alias)
  - *how to select the right entity?*
  - *what to do when disambiguation is not possible?*

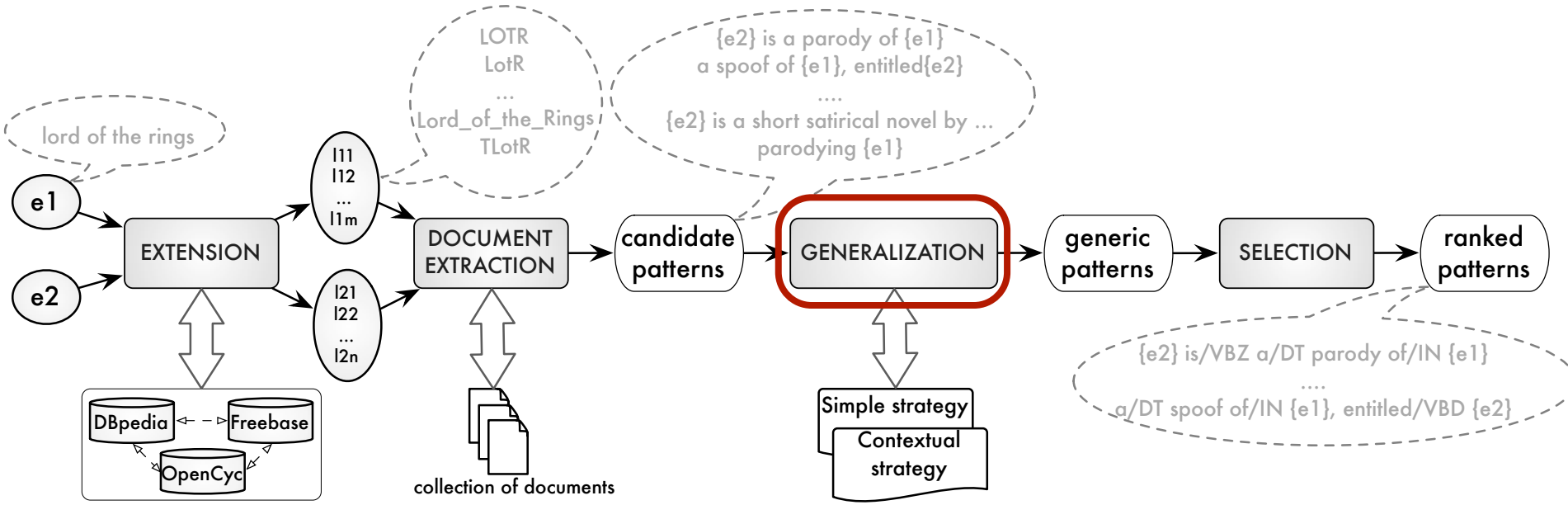
Document extraction



Extracting candidate patterns

- use all the (alternative) labels of entities
- search the collection and rank docs acc. to relevance score
- parse documents, locate sentences with co-occurrence
- consider tokens before and after the entities
- output is a list of candidate patterns

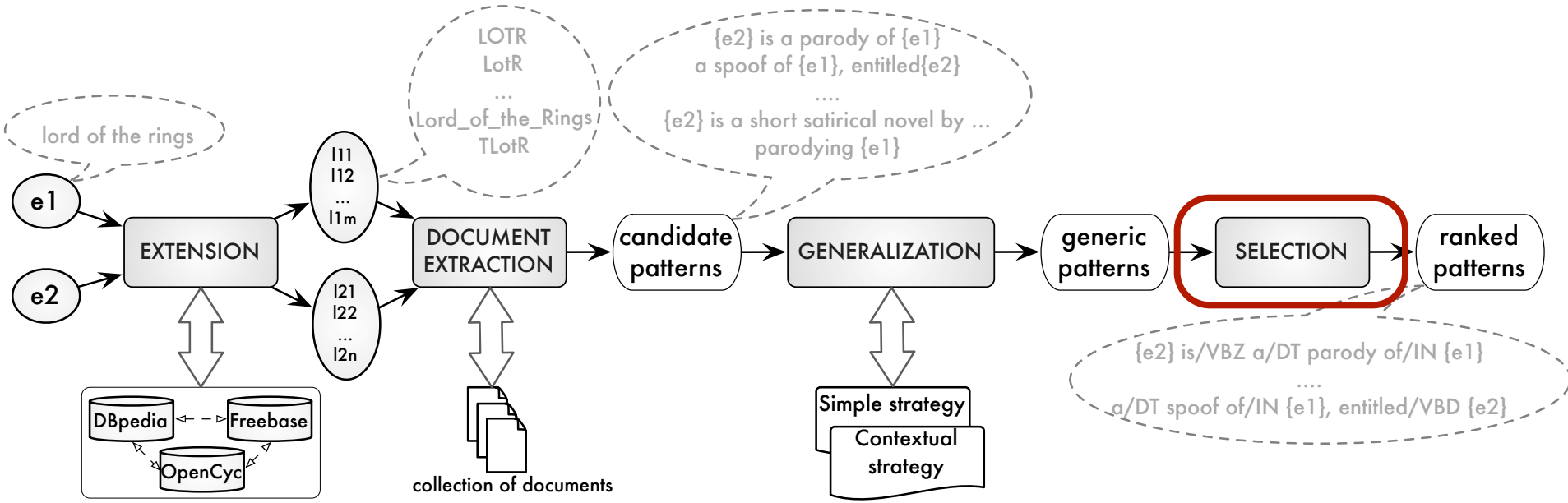
Generalization



Generalization

- goal is to generalize extracted candidate patterns
- use of confidence score to select "best" patterns
- strategies:
  - simple strategy (based on various operations)
    - clean, tag, merge
    - a strategy is sequence of operations
  - contextual strategy based on term frequency
    - most candidate patterns contain a few interesting terms to denote the type of relationship (e.g. Bored of the Rings is a parody of Lord of the Rings)
    - not only terms in between, but also the surrounding context
    - use Wordnet to build a cluster of similar (hyponyms, synonyms) words
    - *e.g. pair "Lord of the Rings" and "Tolkien" leads to "book", "fantasy", "writer" clusters*

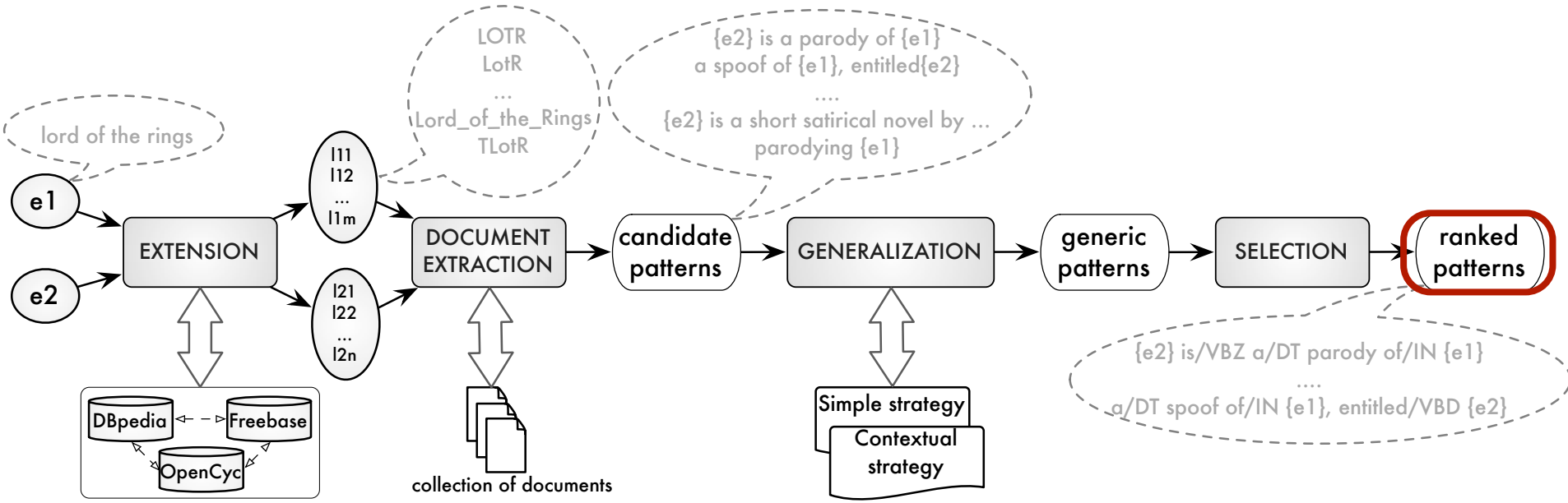
Selection


lord of the rings
e1
e2
EXTENSION
I11
I12
...
I1m
I21
I22
...
I2n
DBpedia
Freebase
OpenCyc
LOTR
LotR
...
Lord_of_the_Rings
TLotR
DOCUMENT EXTRACTION
collection of documents
{e2} is a parody of {e1}
a spoof of {e1}, entitled{e2}
....
{e2} is a short satirical novel by ...
parodying {e1}
candidate patterns
GENERALIZATION
Simple strategy
Contextual strategy
generic patterns
SELECTION
ranked patterns
{e2} is/VBZ a/DT parody of/IN {e1}
....
a/DT spoof of/IN {e1}, entitled/VBD {e2}

Selection

- exploit all information which allowed the discovery of the patterns and to compare patterns
- confidence score: $conf(p) = \left(\frac{\alpha sup_p + \beta occ_p + \gamma prov_p}{\alpha + \beta + \gamma} \right)$
- support: the ratio between the # of examples a pattern is able to discover and the total # examples discovered by all patterns
- occurrency: # of candidate patterns which generalizes a pattern
- provenance: takes into account the document properties in which a pattern was discovered (PageRank, SpamScore, and relevance score)
  - PageRank
  - SpamScore
  - RelScore

Ranked patterns



Relation and Example Discovery

- use patterns to generate examples
- two main use-cases:
  - relationship discovery
  - discover new examples

Relation and Example Discovery (2)

Exploiting patterns

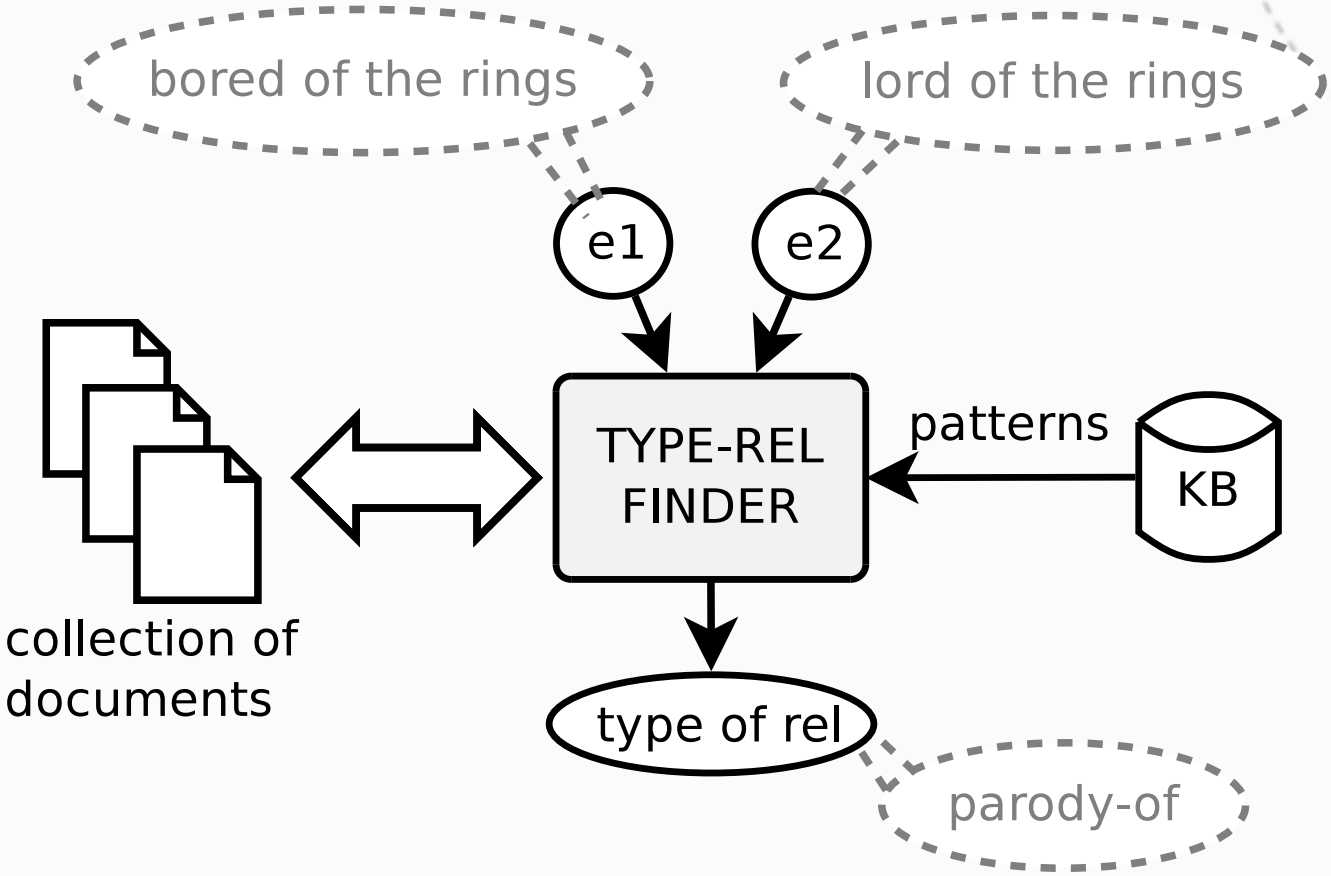
- pattern similarity
  - similarity between a sentence and a pattern
    - presence of frequent terms in a sentence
    - positions of the words
  - calculate the semantic distance

$$semdist(w_p^i, w_s^j) = \begin{cases} 0.0 & \text{if } w_s^j \in \mathcal{FT}_p \\ resnik(w_p^i, w_s^j) & \text{if } w_s^j \notin \mathcal{FT}_p \text{ and } t_p^i = t_s^j \\ 1.0 & \text{otherwise} \end{cases}$$

  - flexible because missing or extra words in the sentence do not affect significantly the similarity score
- Named Entity Recognition (NER)
  - NER is important and an integral part of the approach
  - need to identify and extract entities
  - rely on the formalism of our patterns (since they have been POS-tagged, the tags serve as a delimiter and may constrain the candidate entities)

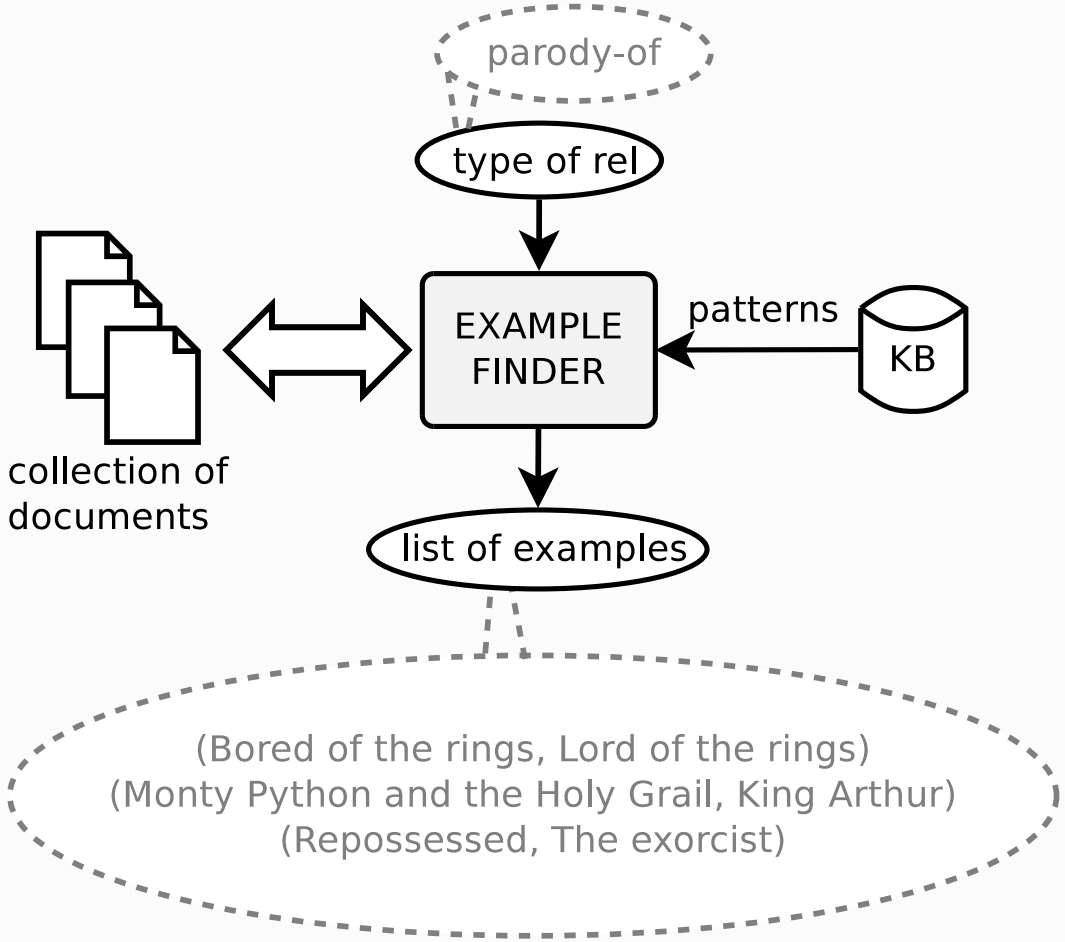
Relation and Example Discovery (3)

Discovering the type of relationship



Relation and Example Discovery (4)

Discovering new examples



Scalability

- distributed processing with MapReduce
- corpus is indexed with Hadoop
- use Pig to compute statistics
- for some tasks, we can partition documents, ordered by the PageRank and /or SpamScore
- process documents by each partition to discover patterns

Experimental study

- English subset of ClueWeb09 collection (~500m documents)
- sentence splitting and tokenization - OpenNLP
- tagging - StanfordNLP
- evalution metrics:
  - precision: number of correct discovered results divided by the total number of discovered results
  - recall: sample-based estimation (3600 manually validated entries by 8 people)
  - F-measure: harmonic mean

Experimental study (2)

Relationship discovery

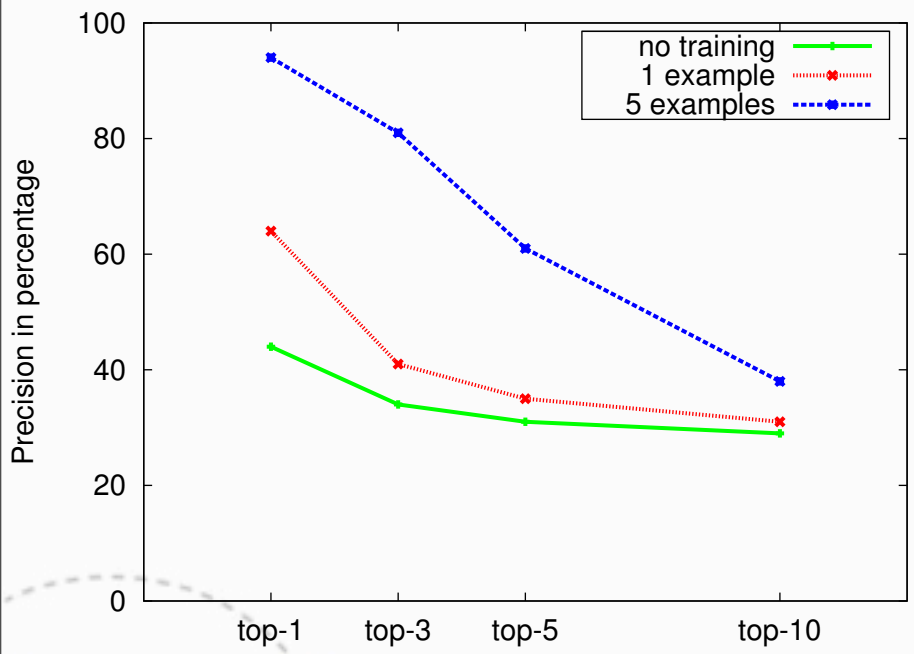
- ground truth
  - manually constructed (200 relationships)
- examples

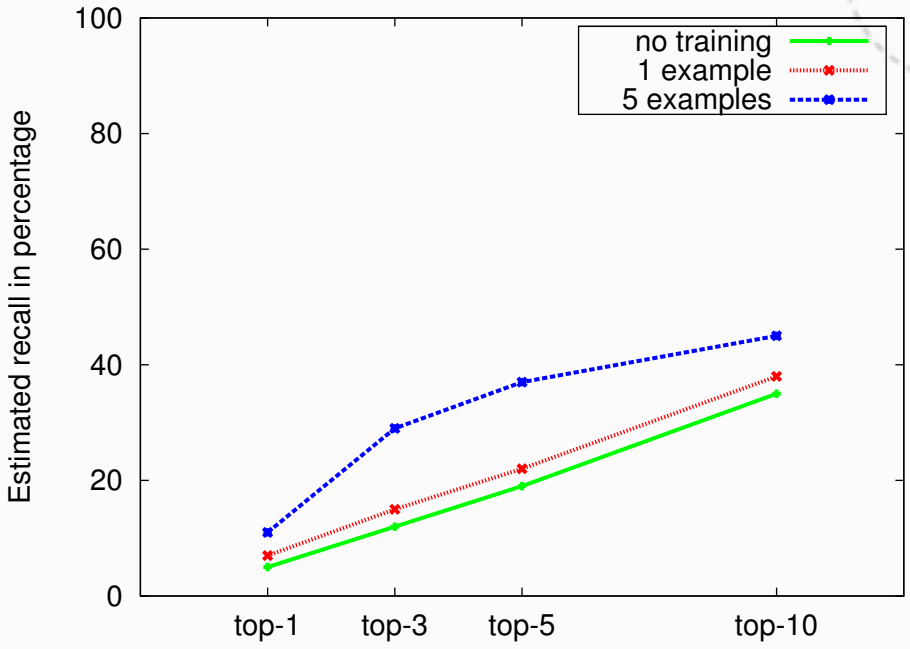
Example	Discovered type of relationship	Confidence score
Obama, Hawai	**birthplace**	0.42
	senator	0.31
	president-elect	0.18
cockatoo, yellow	amazon	0.32
	parrot	0.31
	tail	0.16
eucalyptus, myrtaceae	plant	0.51
	family	0.43
	specie	0.27
Bartolomeo Cristofori, piano	**inventor**	0.60
	instrument	0.43
	maker	0.19
Dave Grohl, Nirvana	Cobain	0.34
	band member	0.27
	drummer	0.16
Bored of the Rings, Lord of the Rings	**parody**	0.53
	links	0.24
	Middle-Earth	0.23

Experimental study (3)

Relationship discovery

- quality is measured by top-k (ranked list of relationship candidates)
- can run both w/ and without training data





Conclusion

- SPIDER - relation extraction at Web-scale
- two-step approach
  - pattern generation
  - relation and example generation
- future work
  - experiments from different domains (recently released dataset - ClueWeb2012)
  - study impact of parameters and contradictory cases
  - enriching instances with attributes
  - open up the interface and integrate the user feedback (GUI, REST API, and SPARQL endpoint)
  - a demo is under submission

Thank you for your attention!

Questions, comments, feedback?

Naimdjon Takhirov
takhirov@idi.ntnu.no