NTNU – Trondheim
Norwegian University of
Science and Technology

Université Claude Bernard Lyon 1

LIRIS

An evidence-based verification approach to extract entities and relations for knowledge base population

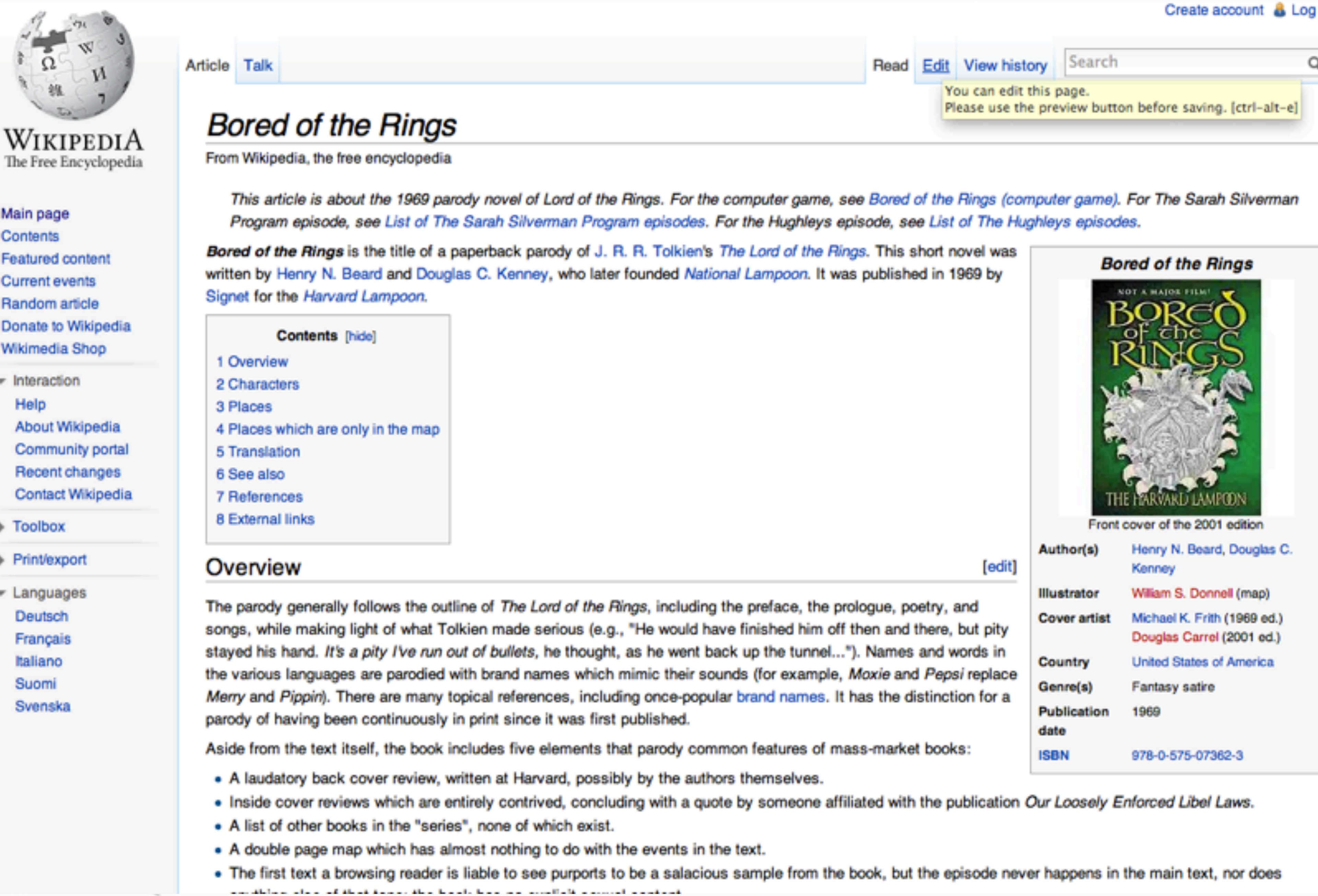
Naimdjon Takhirov, Fabien Duchateau, Trond Aalberg

ISWC'2012 Boston, USA

November 15, 2012

Knowledge extraction

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



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*.

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

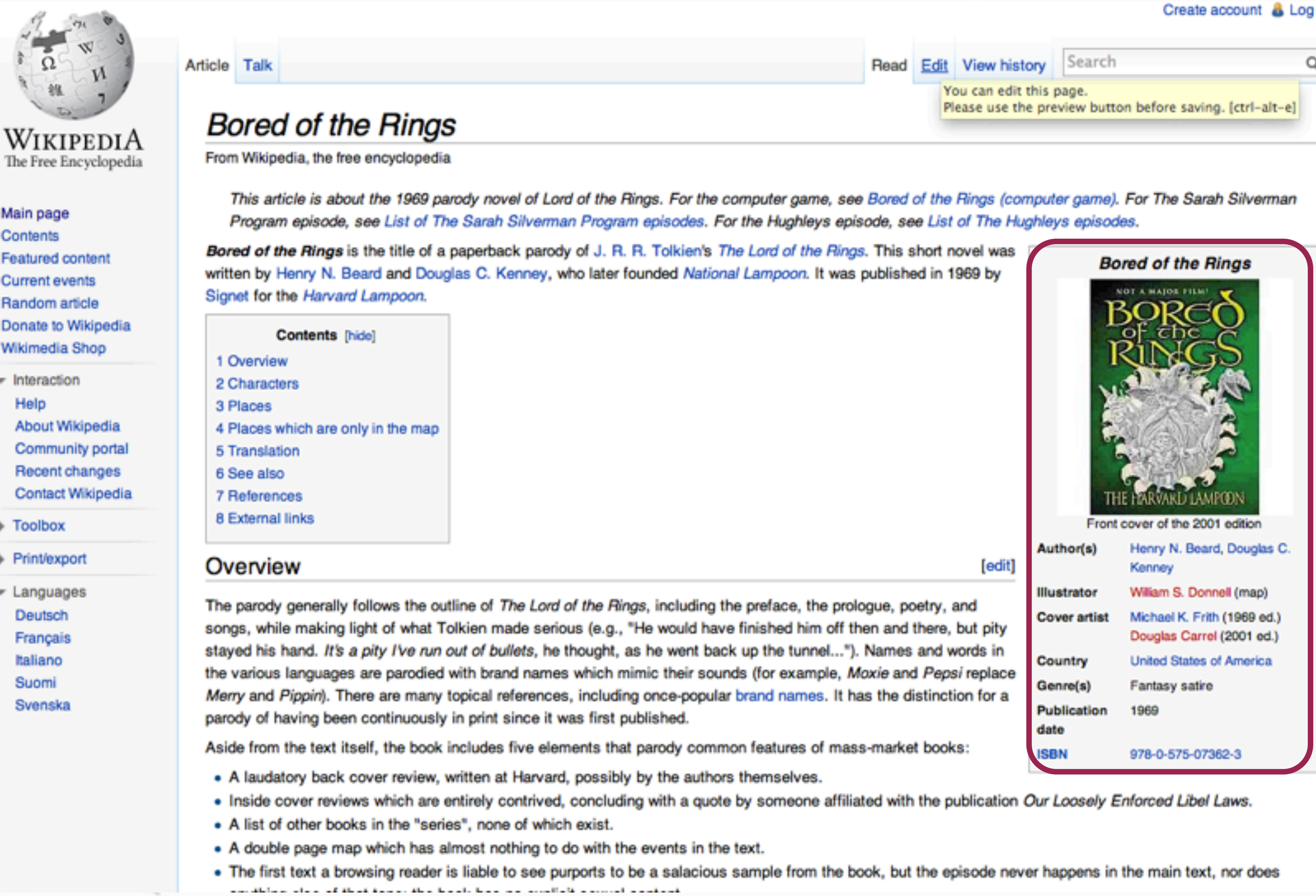
Contents [hide]

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 anything else of that tone; the book has no explicit sexual content.



You can edit this page.
Please use the preview button before saving. [ctrl-alt-e]

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]

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

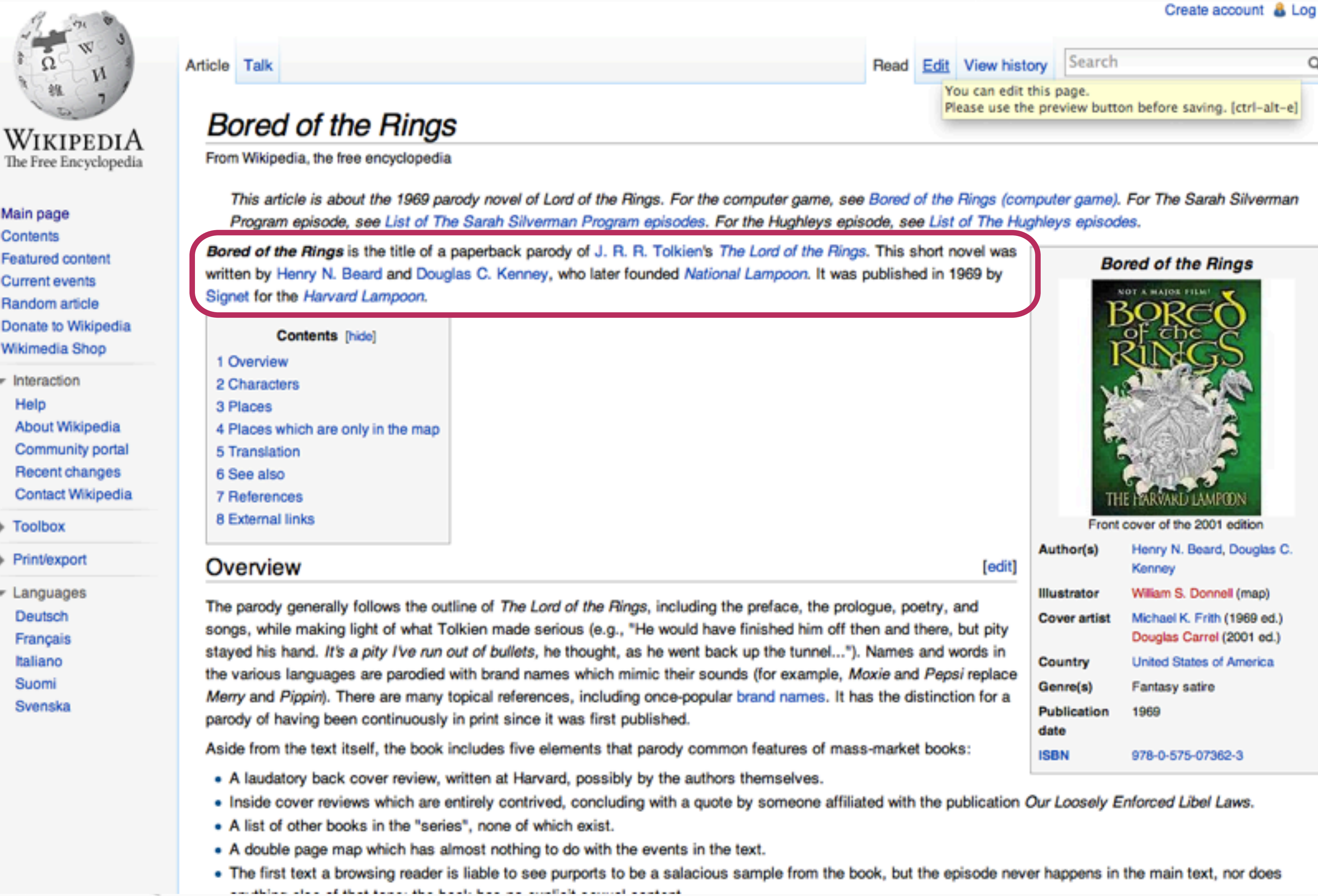
Overview

[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 anything else of that type; the book has no explicit sexual content.



- 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

You can edit this page.
Please use the preview button before saving. [ctrl-alt-e]

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*.

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

Contents [hide]

Overview

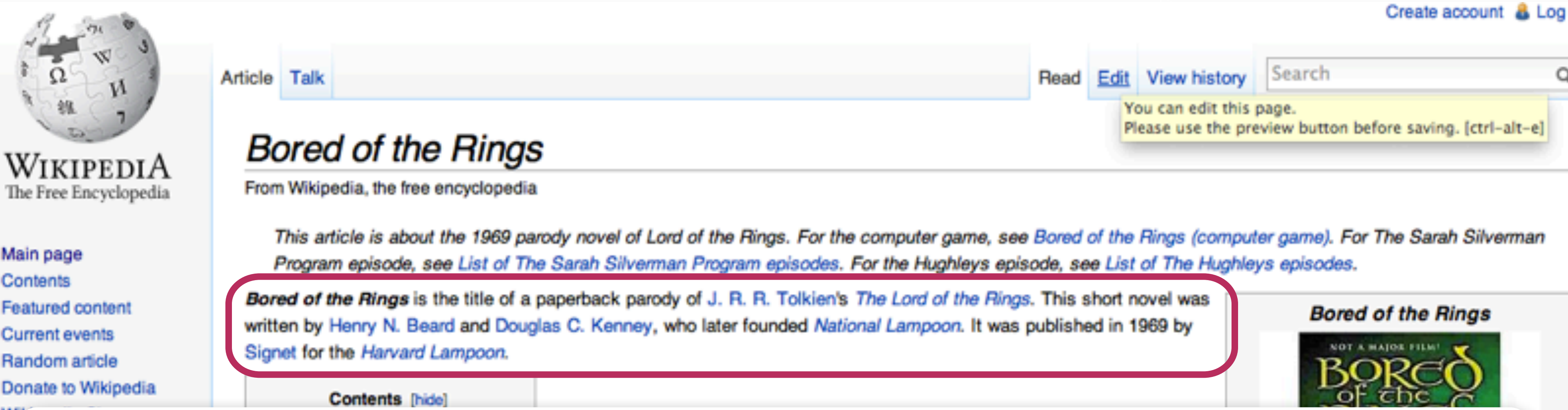
[edit]

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 anything else of that type; the book has no explicit sexual content.

An evidence based verification approach to extract entities and relations for knowledge base population



You can edit this page.
Please use the preview button before saving. [ctrl-alt-e]

Bored of the Rings

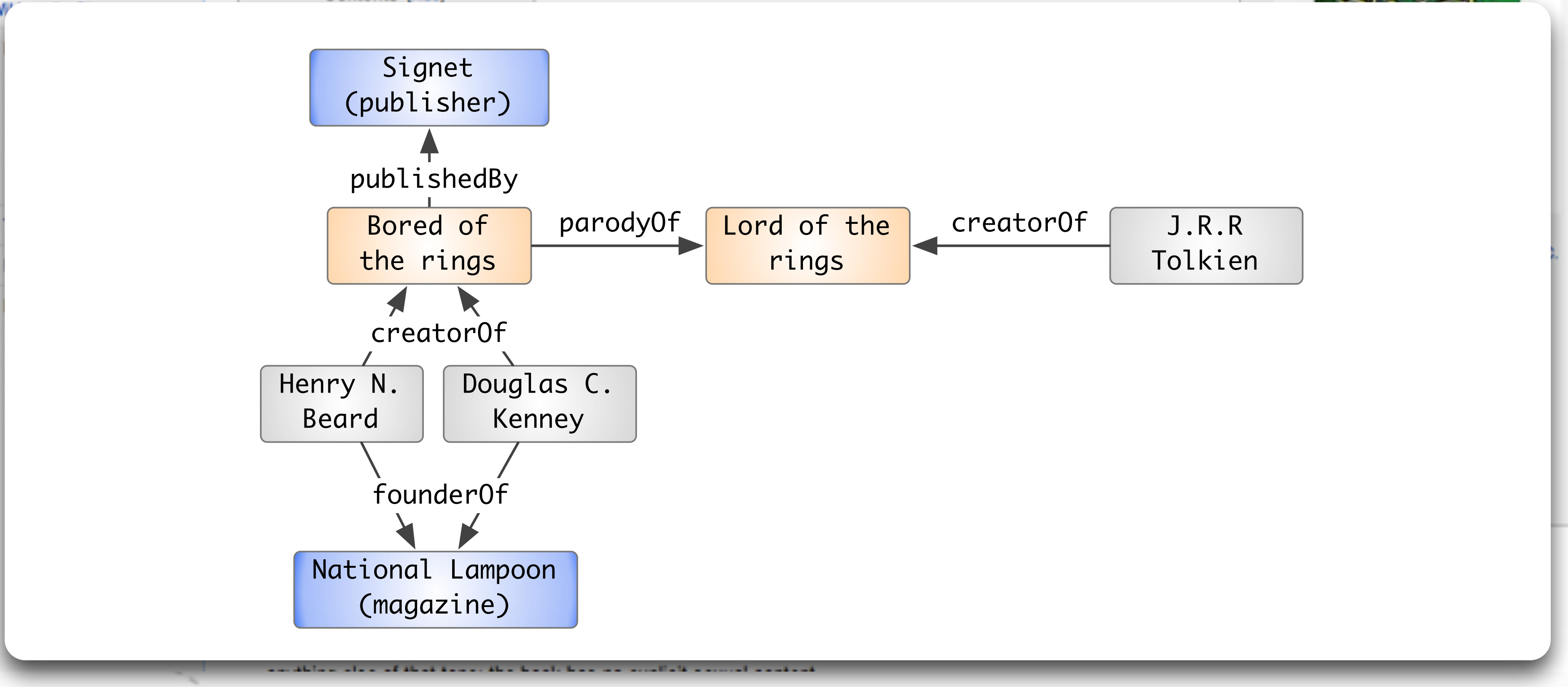
From Wikipedia, the free encyclopedia

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]

Bored of the Rings



An evidence based verification approach to extract entities and relations for knowledge base population

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
- issues: not typed entities/relations, multiple relations, temporal aspect, tradeoff recall/precision, runtime performance

Agenda

- introduction and overview
- approach
  - discovering examples
  - verification
    - classification
    - linking
- experimental evaluation
- conclusion

Introduction

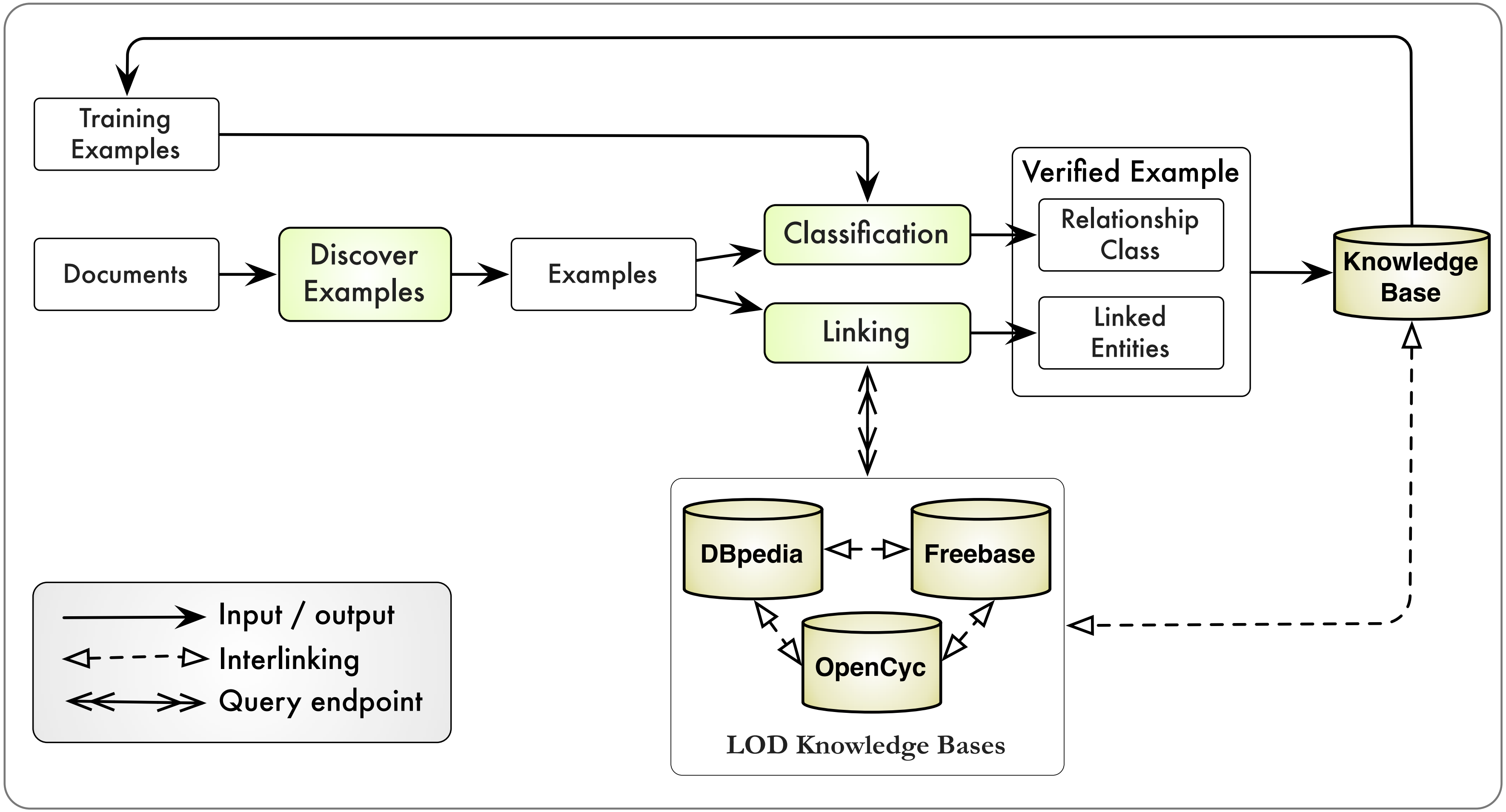
- existing, domain specific data models (e.g. libraries) need an “upgrade”
  - data created several decades ago (legacy data)
  - large investments (on the infrastructure and manpower)
- new semantic data models require a complete conversion
- recent developments of LOD and interest in semantic data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

Introduction (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.)

- KIEV - **K**nowledge and **I**nformation **E**xtraction with **V**erification
  - extracting semantic information from the documents
  - verification with classification and linking techniques
  - reasonable recall/precision wrt state-of-the-art

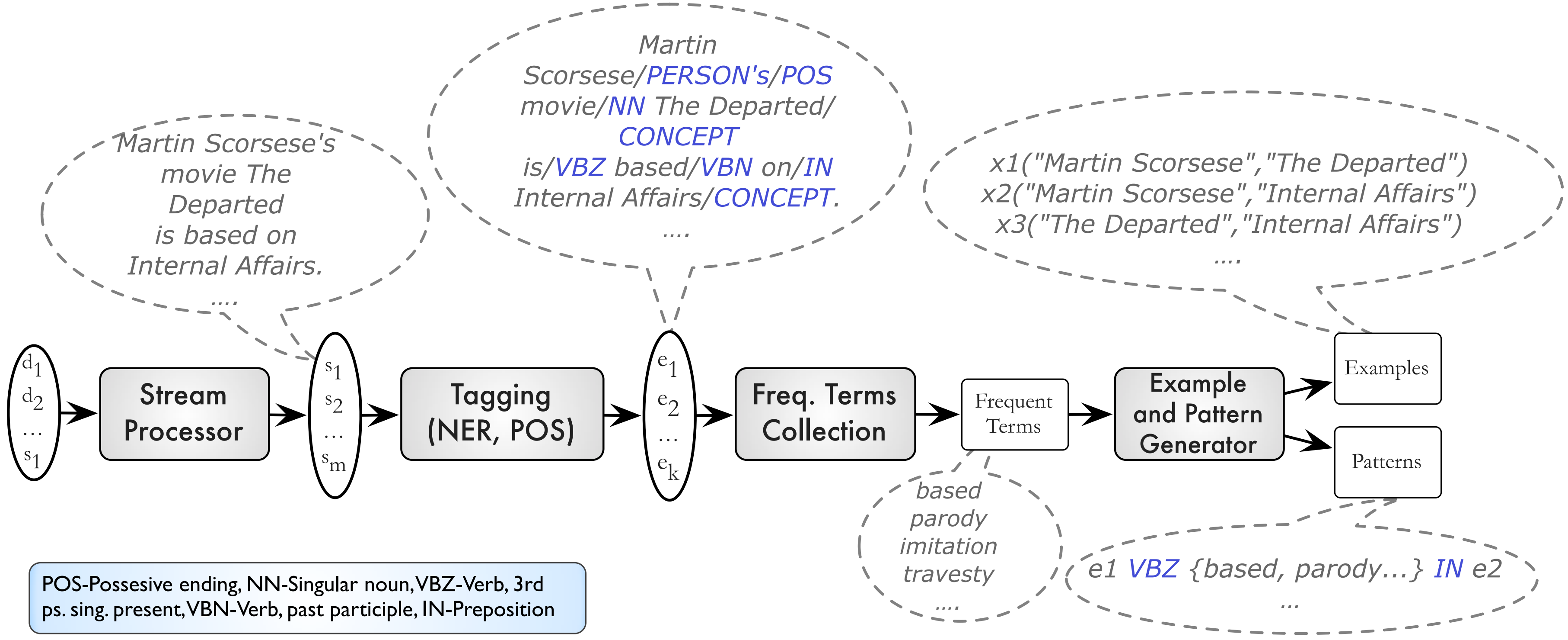
Overview of KIEV



An evidence based verification approach to extract entities and relations for knowledge base population

Discovering examples

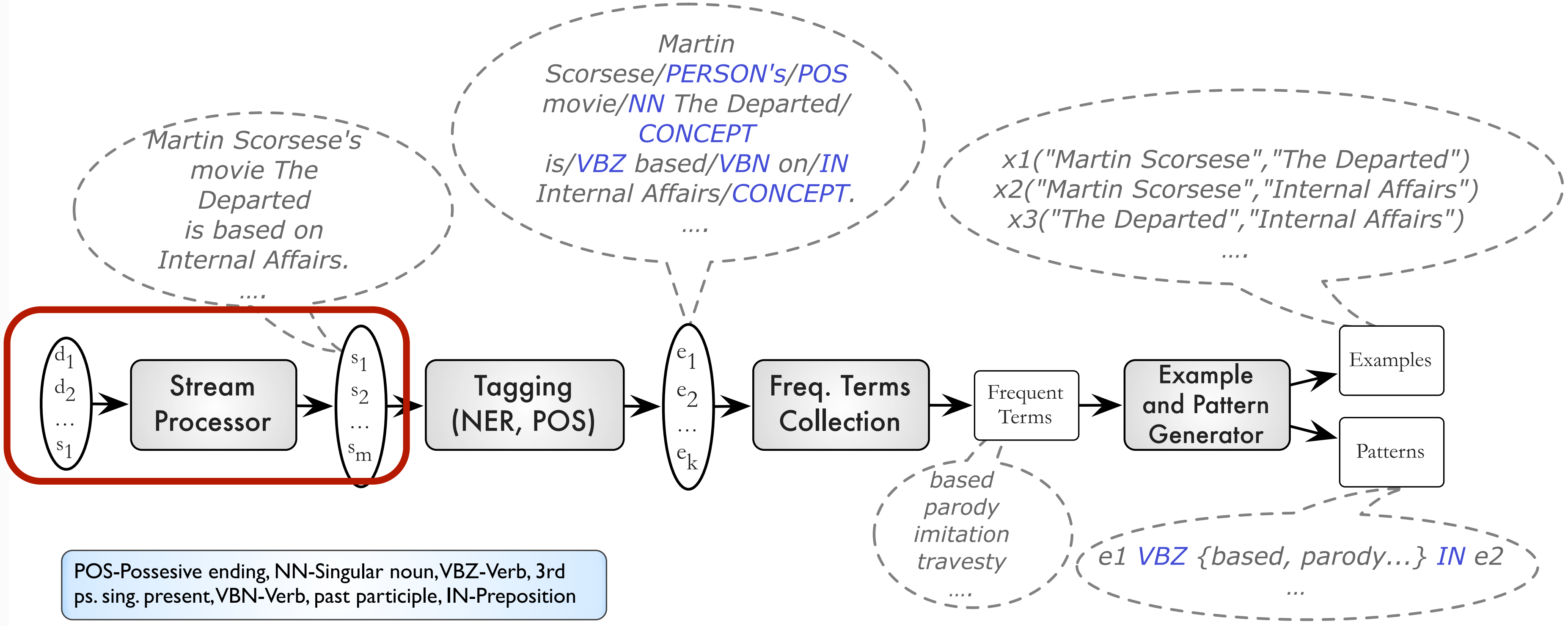
an iterative process to discover relations


d1 d2 … s1
Stream Processor
s1 s2 … sm
Tagging (NER, POS)
e1 e2 … ek
Freq. Terms Collection
Frequent Terms
Example and Pattern Generator
Examples
Patterns
Martin Scorsese's movie The Departed is based on Internal Affairs. ….
Martin Scorsese/PERSON's/POS movie/NN The Departed/ CONCEPT is/VBZ based/VBN on/IN Internal Affairs/CONCEPT. ….
x1("Martin Scorsese","The Departed")
x2("Martin Scorsese","Internal Affairs")
x3("The Departed","Internal Affairs")
….
based parody imitation travesty ….
e1 VBZ {based, parody…} IN e2
…

POS-Possesive ending, NN-Singular noun, VBZ-Verb, 3rd ps. sing. present, VBN-Verb, past participle, IN-Preposition

Discovering examples

an iterative process to discover relations



Martin Scorsese's movie The Departed is based on Internal Affairs. ....

Martin Scorsese/PERSON's/POS movie/NN The Departed/CONCEPT is/VBZ based/VBN on/IN Internal Affairs/CONCEPT. ....

x1("Martin Scorsese","The Departed")
x2("Martin Scorsese","Internal Affairs")
x3("The Departed","Internal Affairs")
....

d_1 d_2 … s_1

Stream Processor

s_1 s_2 … s_m

Tagging (NER, POS)

e_1 e_2 … e_k

Freq. Terms Collection

Frequent Terms

Example and Pattern Generator

Examples

Patterns

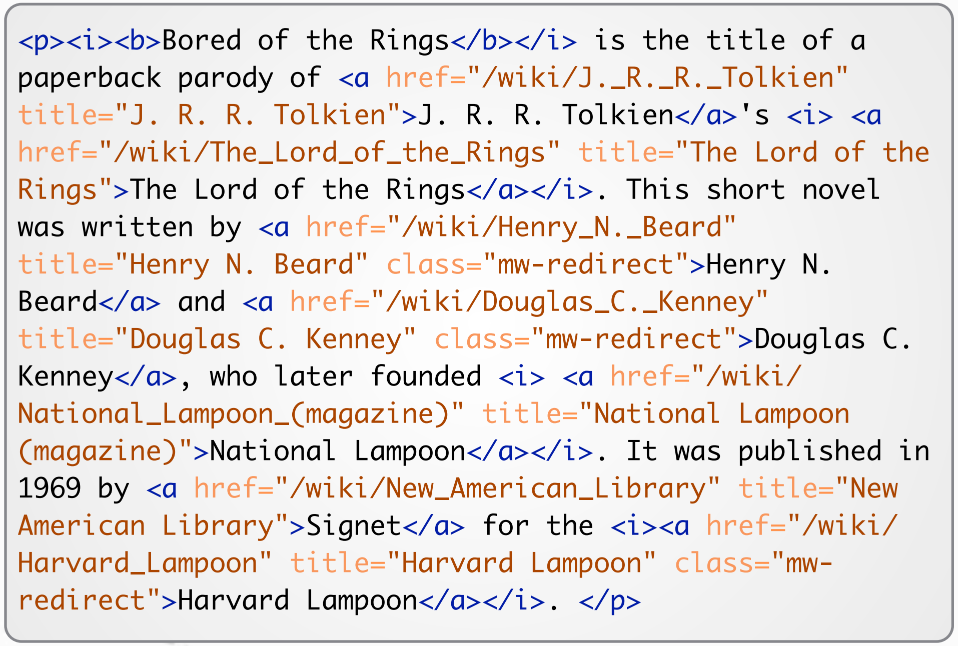
based parody imitation travesty ....

e1 VBZ {based, parody...} IN e2
...

POS-Possesive ending, NN-Singular noun, VBZ-Verb, 3rd ps. sing. present, VBN-Verb, past participle, IN-Preposition

Discovering examples: processing streams

- preprocess the input textual documents
- sentence splitting
- clean noisy sentences



```
<p><i><b>Bored of the Rings</b></i> is the title of a
paperback parody of <a href="/wiki/J._R._R._Tolkien"
title="J. R. R. Tolkien">J. R. R. Tolkien</a>'s <i> <a
href="/wiki/The_Lord_of_the_Rings" title="The Lord of the
Rings">The Lord of the Rings</a></i>. This short novel
was written by <a href="/wiki/Henry_N._Beard"
title="Henry N. Beard" class="mw-redirect">Henry N.
Beard</a> and <a href="/wiki/Douglas_C._Kenney"
title="Douglas C. Kenney" class="mw-redirect">Douglas C.
Kenney</a>, who later founded <i> <a href="/wiki/
National_Lampoon_(magazine)" title="National Lampoon
(magazine)">National Lampoon</a></i>. It was published in
1969 by <a href="/wiki/New_American_Library" title="New
American Library">Signet</a> for the <i><a href="/wiki/
Harvard_Lampoon" title="Harvard Lampoon" class="mw-
redirect">Harvard Lampoon</a></i>. </p>
```

Discovering examples: processing streams

- preprocess the input textual documents
- sentence splitting
- clean noisy sentences

```
<p><i><b>Bored of the Rings</b></i> is the title of a
paperback parody of <a href="/wiki/J._R._R._Tolkien"
title="J. R. R. Tolkien">J. R. R. Tolkien</a>'s <i> <a
href="/wiki/The_Lord_of_the_Rings" title="The Lord of the
Rings">The Lord of the Rings</a></i>. This short novel
was written by <a href="/wiki/Henry_N._Beard"
title="Henry N. Beard" class="mw-redirect">Henry N.
Beard</a> and <a href="/wiki/Douglas_C._Kenney"
title="Douglas C. Kenney" class="mw-redirect">Douglas C.
Kenney</a>, who later founded <i> <a href="/wiki/
National_Lampoon_(magazine)" title="National Lampoon
(magazine)">National Lampoon</a></i>. It was published in
1969 by <a href="/wiki/New_American_Library" title="New
American Library">Signet</a> for the <i><a href="/wiki/
Harvard_Lampoon" title="Harvard Lampoon" class="mw-
redirect">Harvard Lampoon</a></i>. </p>
```

```
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.
```

Discovering examples: processing streams

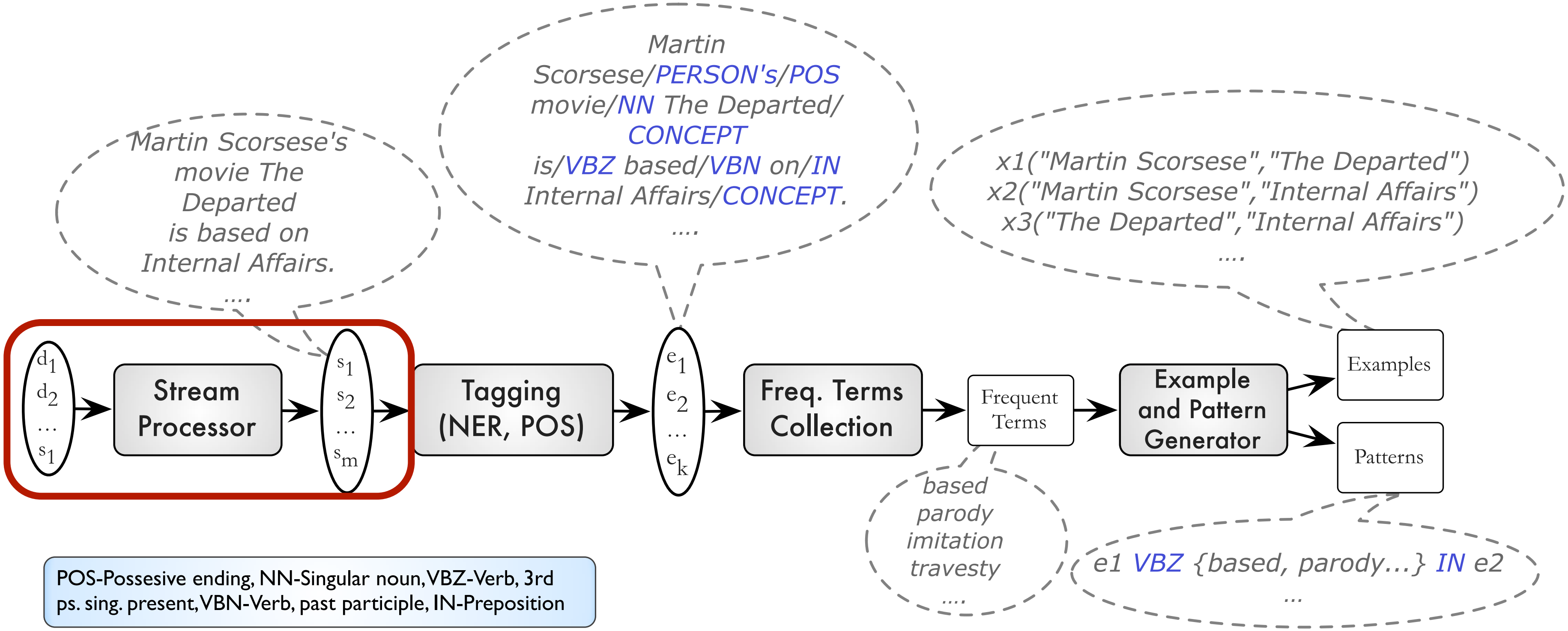
- preprocess the input textual documents
- sentence splitting
- clean noisy sentences

```
<p><i><b>Bored of the Rings</b></i> is the title of a
paperback parody of <a href="/wiki/J._R._R._Tolkien"
title="J. R. R. Tolkien">J. R. R. Tolkien</a>'s <i> <a
href="/wiki/The_Lord_of_the_Rings" title="The Lord of the
Rings">The Lord of the Rings</a></i>. This short novel
was written by <a href="/wiki/Henry_N._Beard"
title="Henry N. Beard" class="mw-redirect">Henry N.
Beard</a> and <a href="/wiki/Douglas_C._Kenney"
title="Douglas C. Kenney" class="mw-redirect">Douglas C.
Kenney</a>, who later founded <i> <a href="/wiki/
National_Lampoon_(magazine)" title="National Lampoon
(magazine)">National Lampoon</a></i>. It was published in
1969 by <a href="/wiki/New_American_Library" title="New
American Library">Signet</a> for the <i><a href="/wiki/
Harvard_Lampoon" title="Harvard Lampoon" class="mw-
redirect">Harvard Lampoon</a></i>. </p>
```

```
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.
```

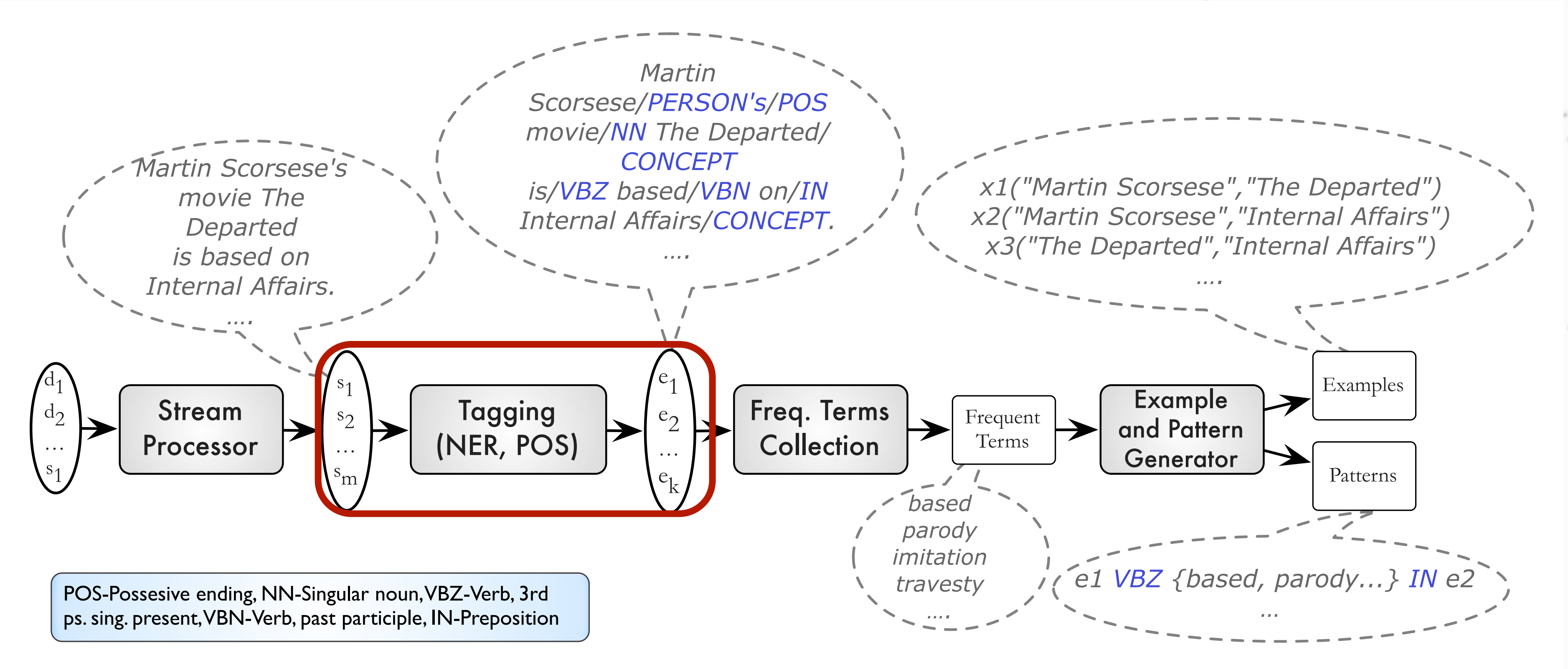
```
d = {
“Bored of the Rings is the title of a paperback
parody of J. R. R. Tolkien's The Lord of the Rings”,

“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.”}
```

Discovering examples: tagging


Martin Scorsese's movie The Departed is based on Internal Affairs.
....
Martin Scorsese/PERSON's/POS movie/NN The Departed/CONCEPT is/VBZ based/VBN on/IN Internal Affairs/CONCEPT.
....
x1("Martin Scorsese","The Departed")
x2("Martin Scorsese","Internal Affairs")
x3("The Departed","Internal Affairs")
....
d_1 d_2 … s_1
Stream Processor
s_1 s_2 … s_m
Tagging (NER, POS)
e_1 e_2 … e_k
Freq. Terms Collection
Frequent Terms
Example and Pattern Generator
Examples
Patterns
based parody imitation travesty
....
e1 VBZ {based, parody...} IN e2
...

POS-Possesive ending, NN-Singular noun, VBZ-Verb, 3rd ps. sing. present, VBN-Verb, past participle, IN-Preposition

Discovering examples: tagging


Martin Scorsese's movie The Departed is based on Internal Affairs. ....
Martin Scorsese/PERSON's/POS movie/NN The Departed/CONCEPT is/VBZ based/VBN on/IN Internal Affairs/CONCEPT. ....
x1("Martin Scorsese","The Departed")
x2("Martin Scorsese","Internal Affairs")
x3("The Departed","Internal Affairs")
....
d_1 d_2 … s_1
Stream Processor
s_1 s_2 … s_m
Tagging (NER, POS)
e_1 e_2 … e_k
Freq. Terms Collection
Frequent Terms
Example and Pattern Generator
Examples
Patterns
based parody imitation travesty ....
e1 VBZ {based, parody…} IN e2
…

POS-Possesive ending, NN-Singular noun, VBZ-Verb, 3rd ps. sing. present, VBN-Verb, past participle, IN-Preposition

Discovering examples: tagging

- identify proper names in text - NER (w/focus on persons, organizations, places and generic concepts)

Discovering examples: tagging

- identify proper names in text - NER (w/focus on persons, organizations, places and generic concepts)

```
Bored_of_the_Rings/CONCEPT is the title of a paperback  parody of
J._R._R._Tolkien/PERSON's The_Lord_of_the_Rings/CONCEPT.
```

Discovering examples: tagging

- identify proper names in text - NER (w/focus on persons, organizations, places and generic concepts)

```
Bored_of_the_Rings/CONCEPT is the title of a paperback  parody of
J._R._R._Tolkien/PERSON's The_Lord_of_the_Rings/CONCEPT.
```

- POS tagging

Discovering examples: tagging

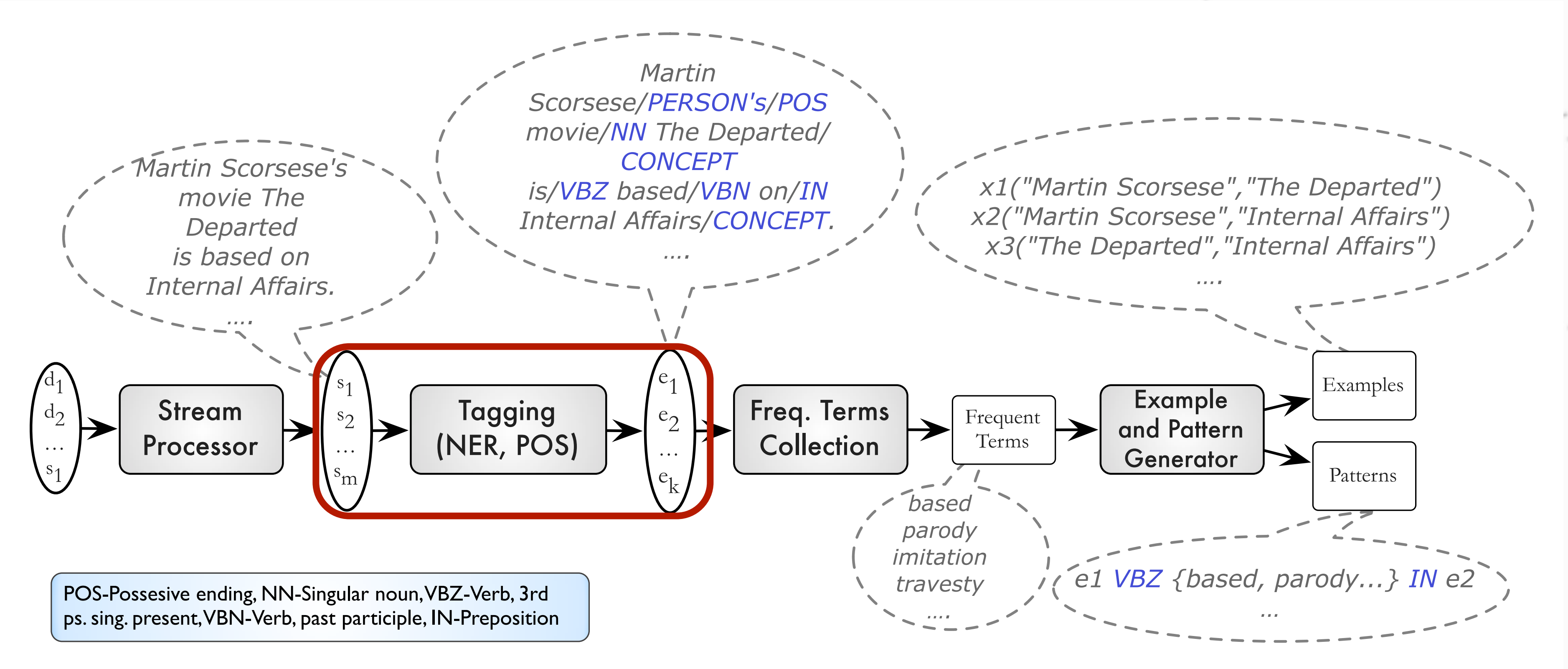
- identify proper names in text - NER (w/focus on persons, organizations, places and generic concepts)

```
Bored_of_the_Rings/CONCEPT is the title of a paperback  parody of
J._R._R._Tolkien/PERSON's The_Lord_of_the_Rings/CONCEPT.
```

- POS tagging

```
Bored_of_the_Rings/CONCEPT is/VBZ the/DT title/NN of/IN a/DT
paperback/NN parody/NN of/IN J._R._R._Tolkien/PERSON's/POS
The_Lord_of_the_Rings/CONCEPT.
```

Discovering examples: frequent terms collection



d_1 d_2 … s_1

Stream Processor

Martin Scorsese's movie The Departed is based on Internal Affairs. ….

s_1 s_2 … s_m

Tagging (NER, POS)

Martin Scorsese/PERSON's/POS movie/NN The Departed/CONCEPT is/VBZ based/VBN on/IN Internal Affairs/CONCEPT. ….

e_1 e_2 … e_k

Freq. Terms Collection

Frequent Terms

based parody imitation travesty ….

Example and Pattern Generator

x1("Martin Scorsese","The Departed") x2("Martin Scorsese","Internal Affairs") x3("The Departed","Internal Affairs") ….

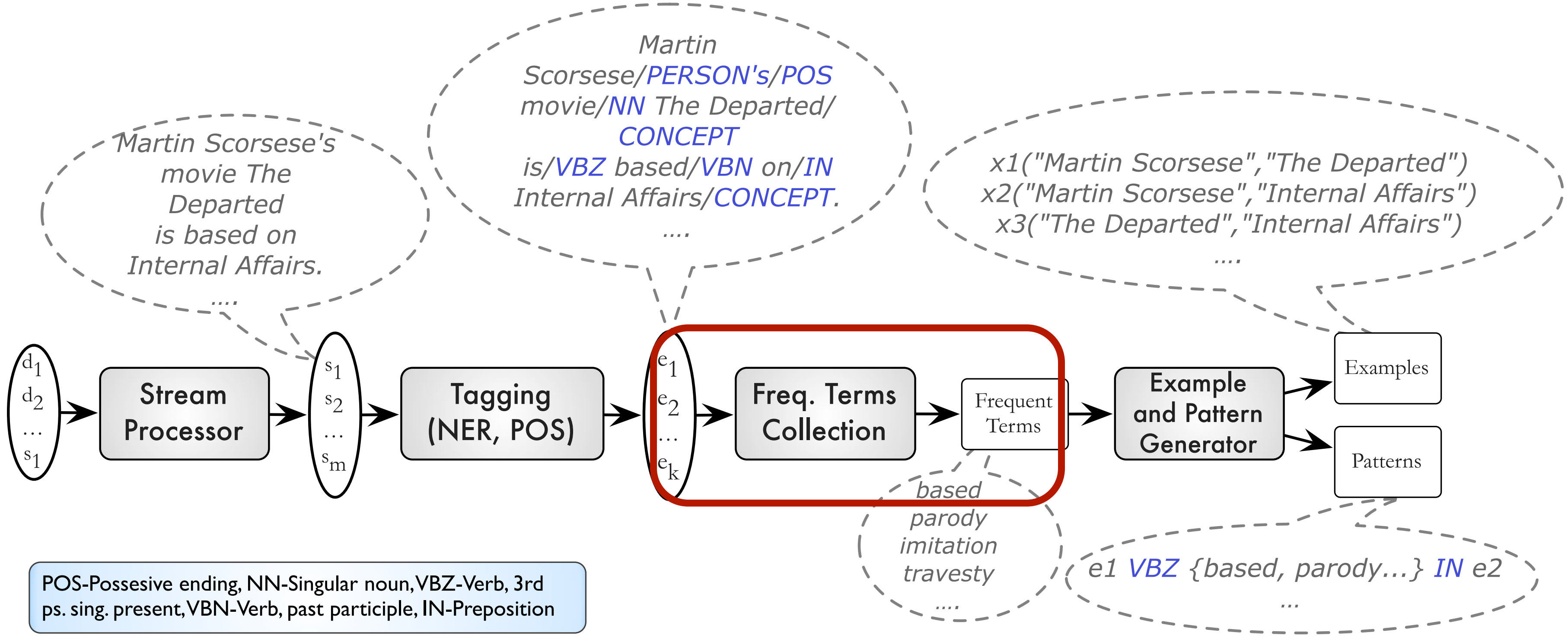
Examples

Patterns

e1 VBZ {based, parody…} IN e2 …

POS-Possesive ending, NN-Singular noun, VBZ-Verb, 3rd ps. sing. present, VBN-Verb, past participle, IN-Preposition

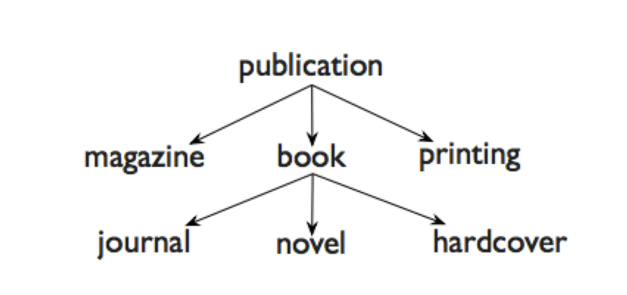
Discovering examples: frequent terms collection


d1
d2
…
s1
Stream Processor
s1
s2
…
sm
Tagging (NER, POS)
e1
e2
…
ek
Freq. Terms Collection
Frequent Terms
Example and Pattern Generator
Examples
Patterns
Martin Scorsese's movie The Departed is based on Internal Affairs.
….
Martin Scorsese/PERSON's/POS movie/NN The Departed/CONCEPT is/VBZ based/VBN on/IN Internal Affairs/CONCEPT.
….
x1("Martin Scorsese","The Departed")
x2("Martin Scorsese","Internal Affairs")
x3("The Departed","Internal Affairs")
….
based
parody
imitation
travesty
….
e1 VBZ {based, parody…} IN e2
…

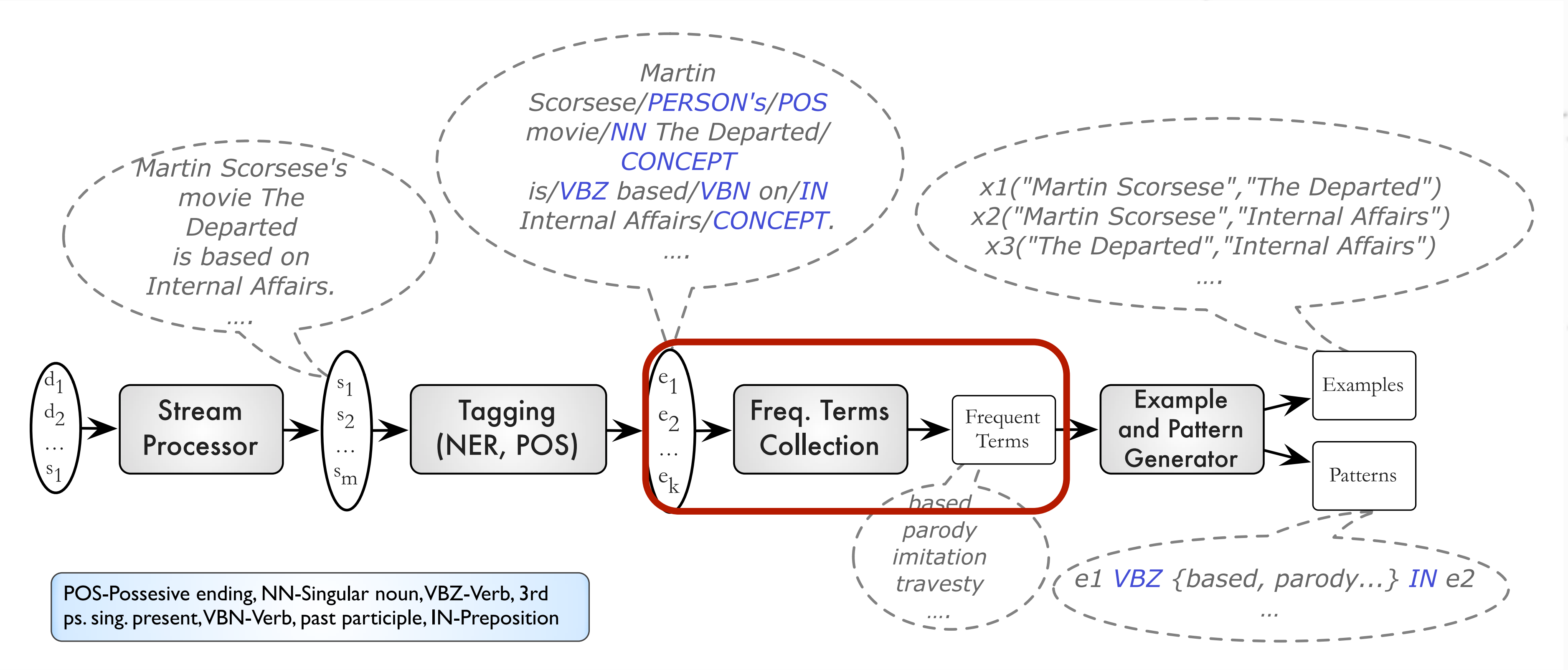
POS-Possesive ending, NN-Singular noun, VBZ-Verb, 3rd ps. sing. present, VBN-Verb, past participle, IN-Preposition

Discovering examples: frequent terms collection

- intuition: terms co-occurring often with the pair of named entities are likely to be relevant
- collect n-grams (exclude common words, e.g.: "and", "the", "of", "a" etc)
- lookup wordnet to obtain the list of semantically related words
- control the granularity of relationships with Resnik taxonomic similarity
  - wordnet synsets (synonyms, hyponyms) : "`writer => novelist`", "`parody => imitation`"
  - `e.g.: Resnik distance between "novel" and "book" is 0.29.`

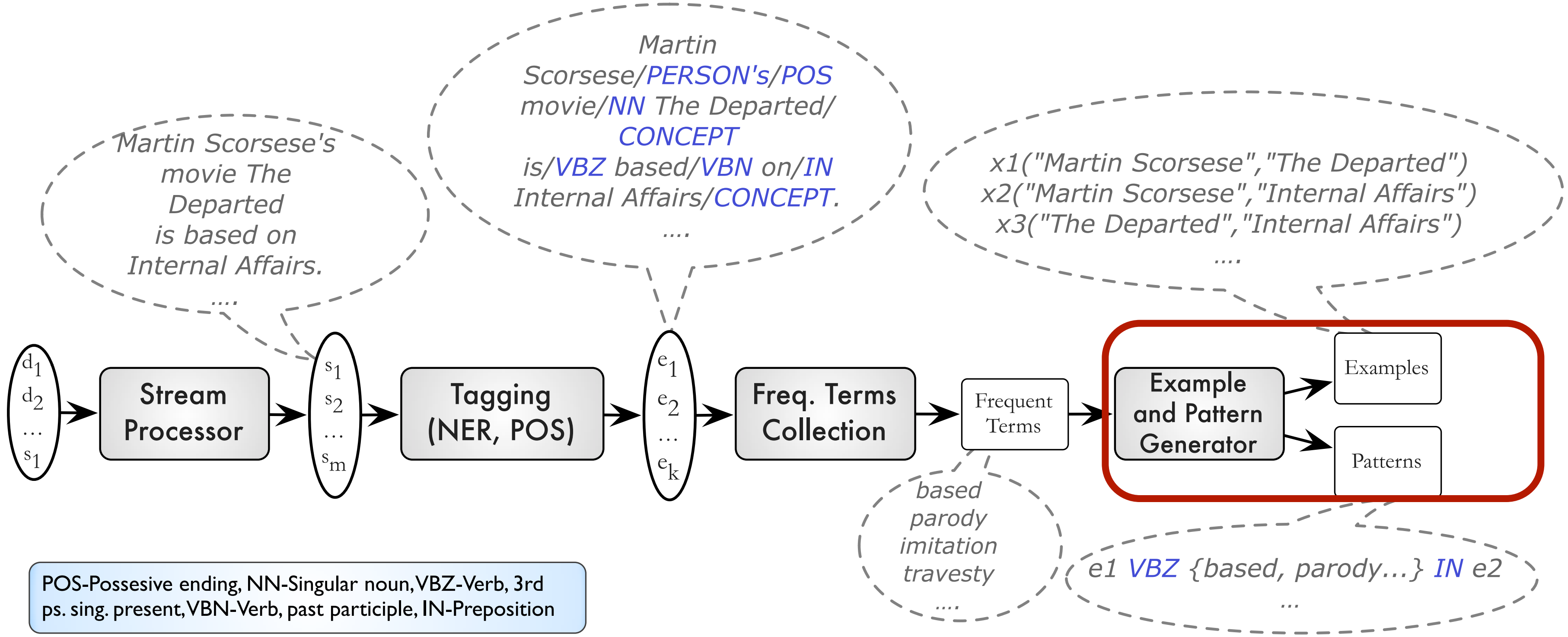


Discovering examples: example & pattern generation


d1
d2
…
s1
Stream Processor
Martin Scorsese's movie The Departed is based on Internal Affairs.
….
s1
s2
…
sm
Tagging (NER, POS)
Martin Scorsese/PERSON's/POS movie/NN The Departed/CONCEPT is/VBZ based/VBN on/IN Internal Affairs/CONCEPT.
….
e1
e2
…
ek
Freq. Terms Collection
Frequent Terms
based
parody
imitation
travesty
….
Example and Pattern Generator
x1("Martin Scorsese","The Departed")
x2("Martin Scorsese","Internal Affairs")
x3("The Departed","Internal Affairs")
….
Examples
Patterns
e1 VBZ {based, parody…} IN e2
…

POS-Possesive ending, NN-Singular noun, VBZ-Verb, 3rd ps. sing. present, VBN-Verb, past participle, IN-Preposition

Discovering examples: example & pattern generation


Martin Scorsese's movie The Departed is based on Internal Affairs.
....
Martin Scorsese/PERSON's/POS movie/NN The Departed/CONCEPT is/VBZ based/VBN on/IN Internal Affairs/CONCEPT.
....
x1("Martin Scorsese","The Departed")
x2("Martin Scorsese","Internal Affairs")
x3("The Departed","Internal Affairs")
....
d1 d2 … s1
Stream Processor
s1 s2 … sm
Tagging (NER, POS)
e1 e2 … ek
Freq. Terms Collection
Frequent Terms
Example and Pattern Generator
Examples
Patterns
based parody imitation travesty ....
e1 VBZ {based, parody…} IN e2
…

POS-Possesive ending, NN-Singular noun, VBZ-Verb, 3rd ps. sing. present, VBN-Verb, past participle, IN-Preposition

Discovering examples: example & pattern generation

- build candidate examples between concepts, persons, etc

```
Bored_of_the_Rings/CONCEPT is/VBZ the/DT title/NN of/IN a/DT
paperback/NN parody/NN of/IN J._R._R._Tolkien/PERSON's/POS
The_Lord_of_the_Rings/CONCEPT.
```

- examples:
  - {Bored_of_the_Rings, J._R._R._Tolkien}
  - {Bored_of_the_Rings, Lord_of_the_Rings}
  - {Lord_of_the_Rings, J._R._R._Tolkien}

Discovering examples: example & pattern generation

- obtain patterns based on the generated examples

```
Bored_of_the_Rings/CONCEPT is/VBZ the/DT title/NN of/IN a/DT
paperback/NN parody/NN of/IN J._R._R._Tolkien/PERSON's/POS
The_Lord_of_the_Rings/CONCEPT.
```

- reuse patterns at the next iteration
- example of patterns extracted from the above sentence:
  - `{e1 is title of e2}`
  - `{e1 paperback parody of e2}`
  - `{e1 /POS e2}`
- ranking patterns $score(p) = \dfrac{\alpha \dfrac{occ(p)}{i} + \beta \dfrac{|P_p|}{|P_i|} + \gamma \dfrac{|X_p|}{|X|}}{\alpha + \beta + \gamma}$

i= no iterations, occ(p)= no of iter. this pattern discovered, Pp=list of merged patterns, Pi=list of generated patterns in this iteration, Xp=support examples, X=total no of examples

Classification

- the task of identifying a relation => classification problem
- given a set of features (properties), the idea is to find the correct class for a given example (extracted from a sentence)
- each class represents a type of relationship (e.g., imitation, creatorOf), e.g.:
    - `{Bored_of_the_Rings, Lord_of_the_Rings} => parodyOf`
    - `{James_Cameron, Avatar} => creatorOf`
- features: document-based, sentence based and entity-based
- two challenges: (i) the choice of training data, (ii) selection of the classifier

Classification (2)

- the choice of training data
  - to improve robustness of the classifier, we need to use more examples as training data after each iteration (incl. bad examples)
  - two strategies to select new training examples
    - linking: select all examples that have been verified at previous iteration
    - frequency: examples discovered in half of the previous iterations
- selection of classifier
  - support flexible classifier selection because the performance of classifiers at various iteration might be of variable quality
  - generate different types of classifiers: decision trees (J48, RandomForest) , instance-based (KStar, IBk), rule-based (NNge, JRip)
  - 10-fold cross-validation
  - the classifier which best minimizes the misclassification rate is selected for each iteration

Linking

- the task of checking/verifying entities of an example => discover corresponding entities on LOD
- query descriptive texts of knowledge bases ("dbpedia-owl:abstract") to obtain initial set of candidates
- use context to construct term vectors of local entity and each candidate LOD entity
  - vector of local entity: tf-idf applied to all documents where entity is mentioned
  - vector of LOD entity: tf-idf applied to its descriptive text
- compute cosine similarity between the vectors and create a link if the similarity is greater than a threshold
- perform extra check on labels to ensure a "reasonable" similarity (w/measures Jaro Winkler, Monge Elkan and Scaled Levenshtein)

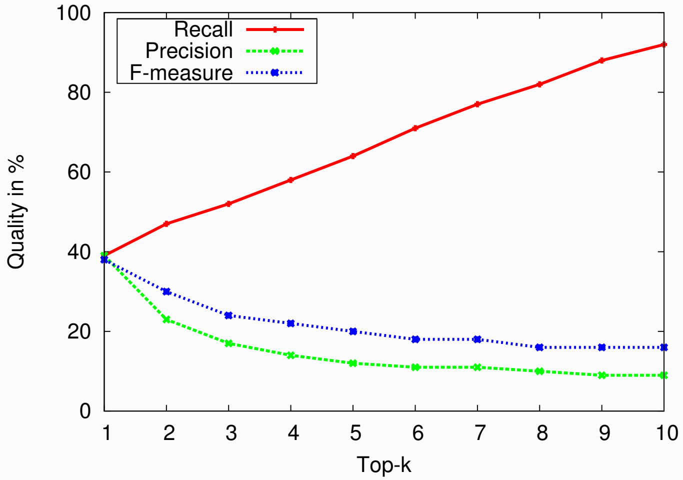
Experimental study

- English subset of ClueWeb09 collection (~500m documents)
- sentence splitting and tokenization - OpenNLP
- tagging - StanfordNLP
- NER - Alchemy API, Zemantha, StanfordNLP
- classification (Weka): Naive Bayes, rule-based (NNge, DecisionTable), tree-based (J48, RandomForest) and lazy (KStar)
- linking - DBpedia v3.7

Dataset and ground truth

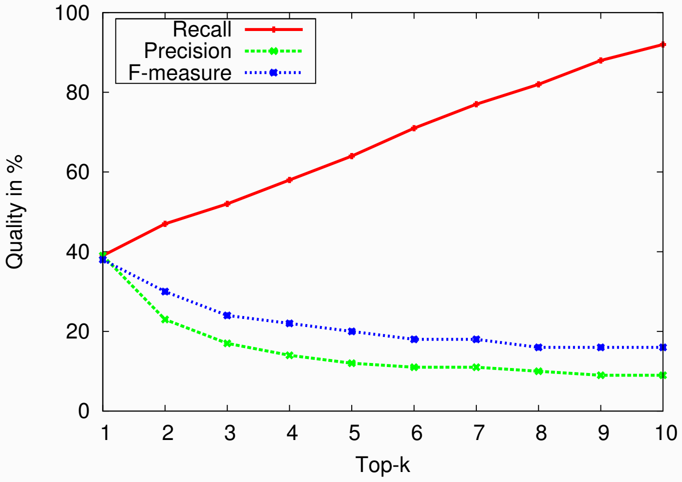
- movie dataset (remakes)
- relations: parodyOf, adaptationOf, creatorOf
- IMDB as ground truth
- 545 remake pairs out of 1052
  - remaining 507 did not have “supporting” documents
    - e.g. movies produced before 70s

Quality of discovery

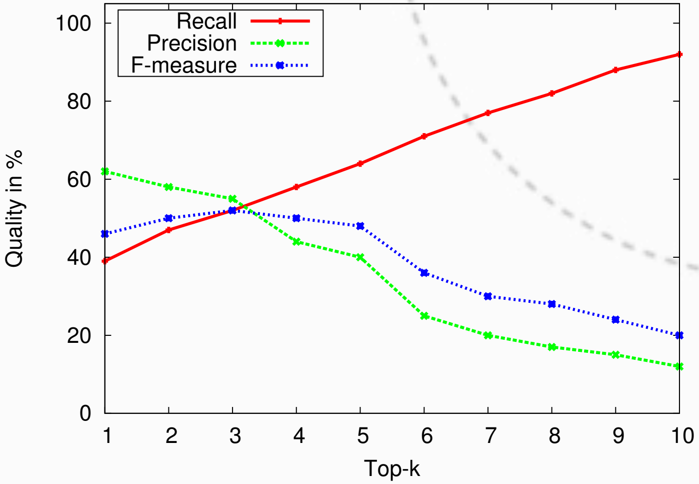


no verification

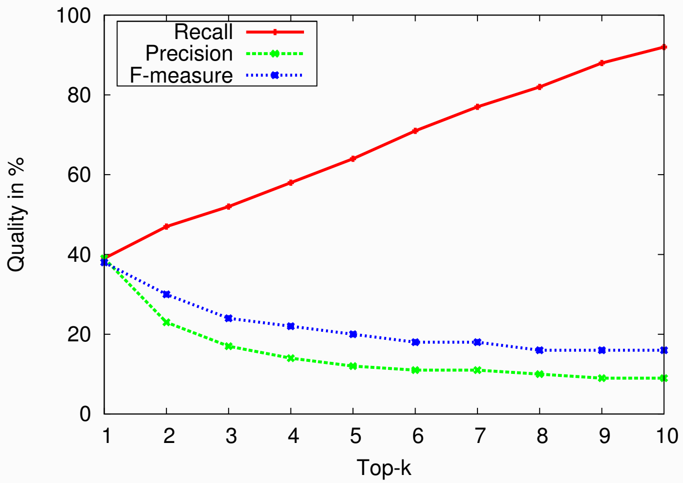
Quality of discovery



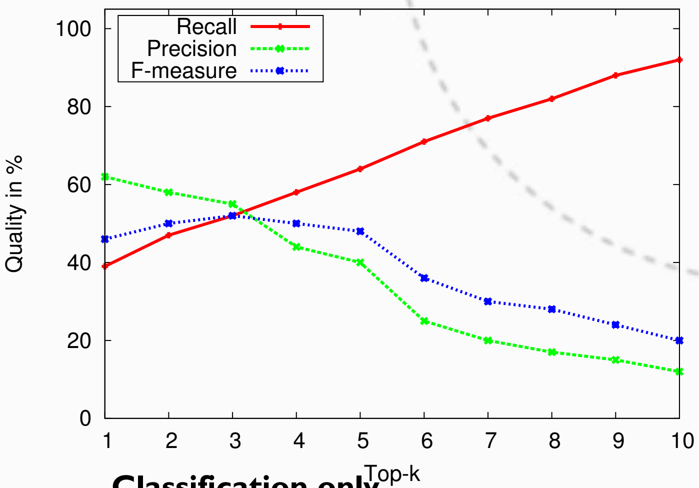
no verification



Quality of discovery

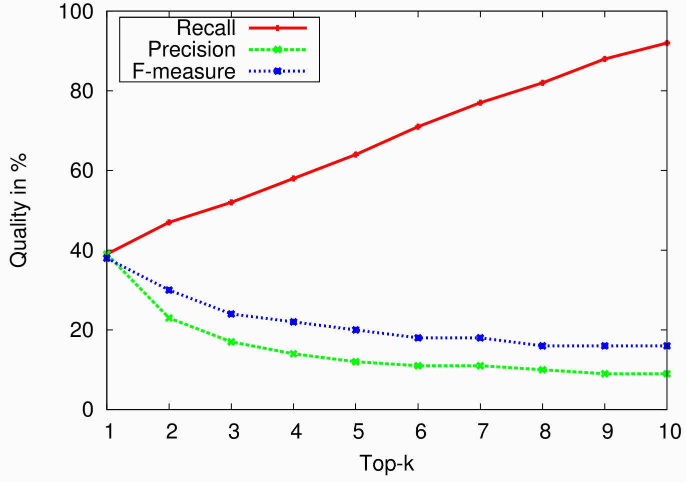


no verification

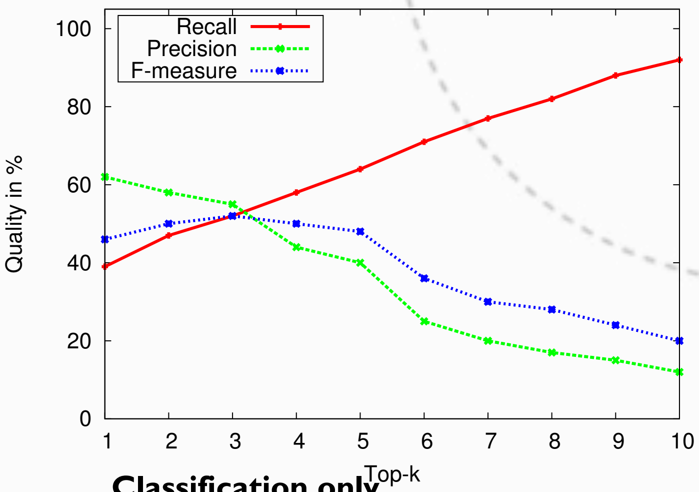


Classification only

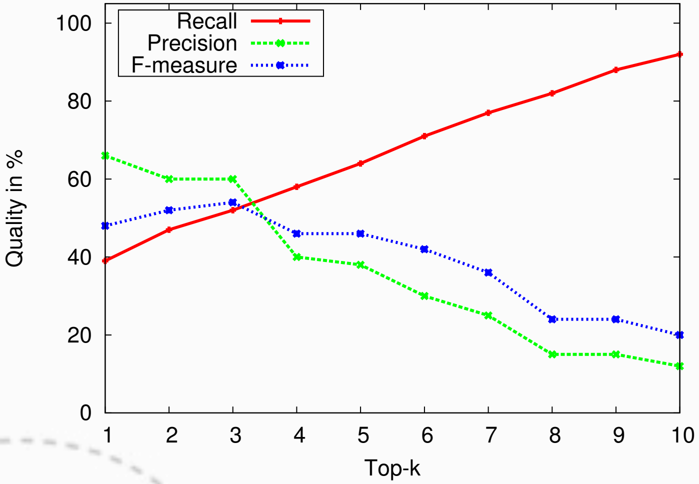
Quality of discovery



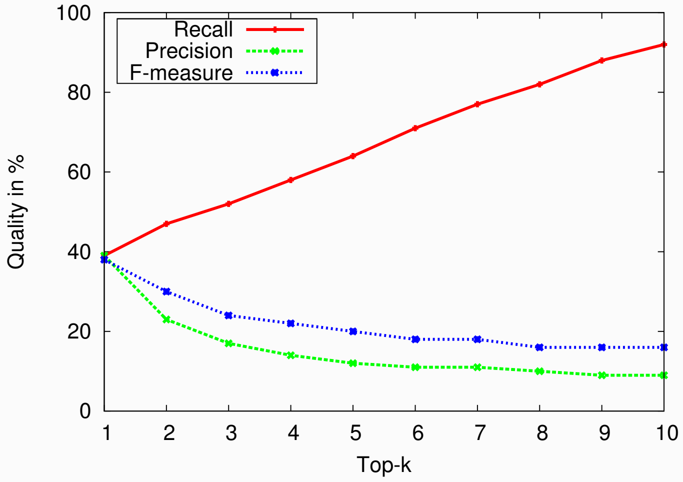
no verification



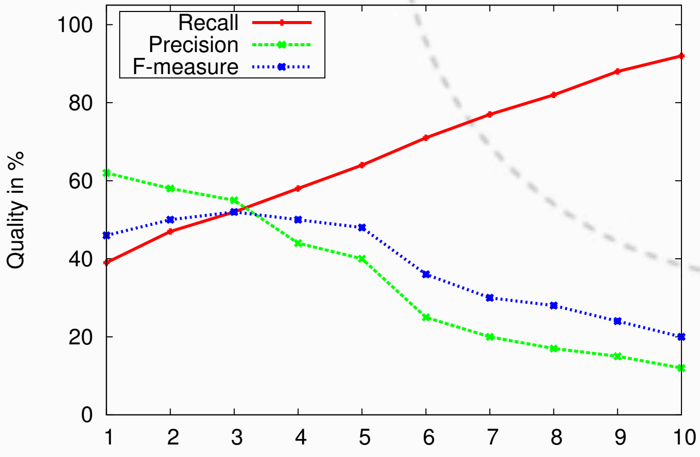
Classification only



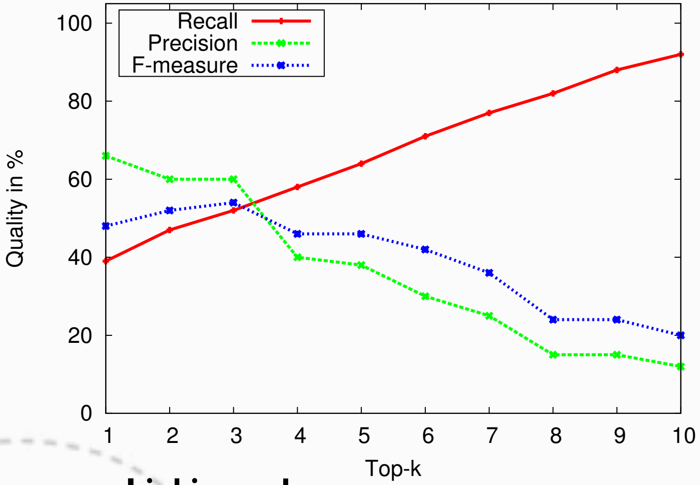
Quality of discovery



no verification

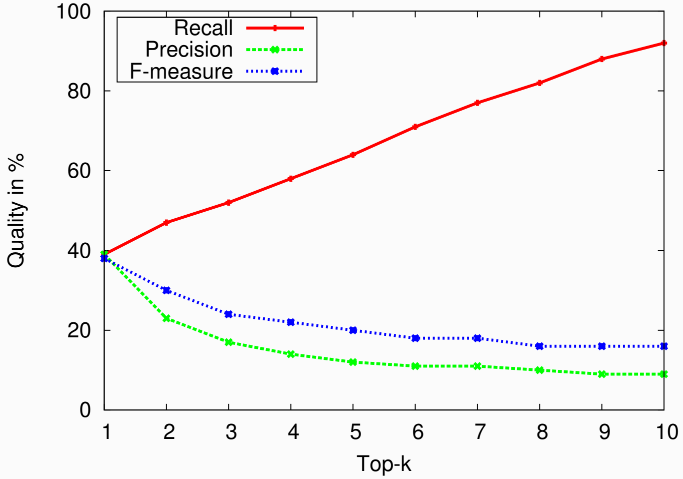


Classification only

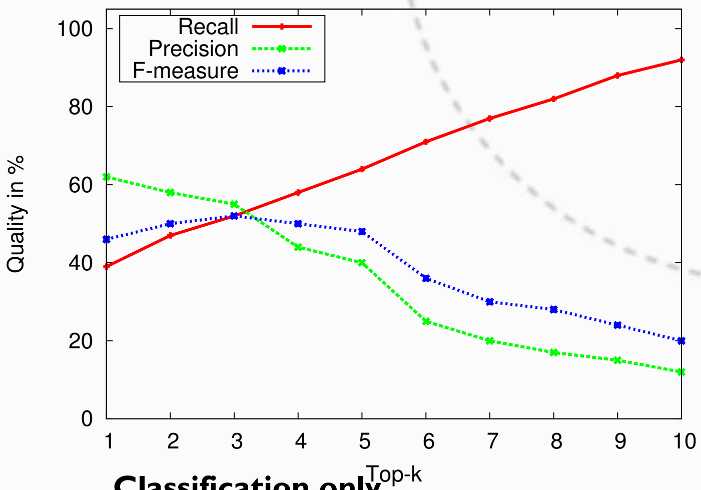


Linking only

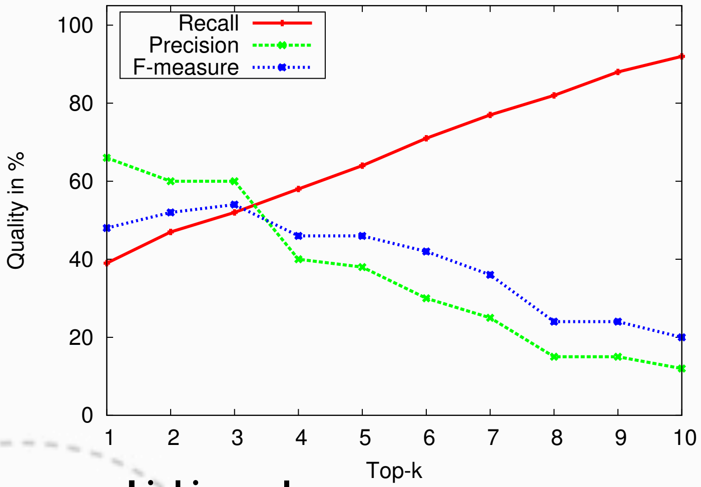
Quality of discovery



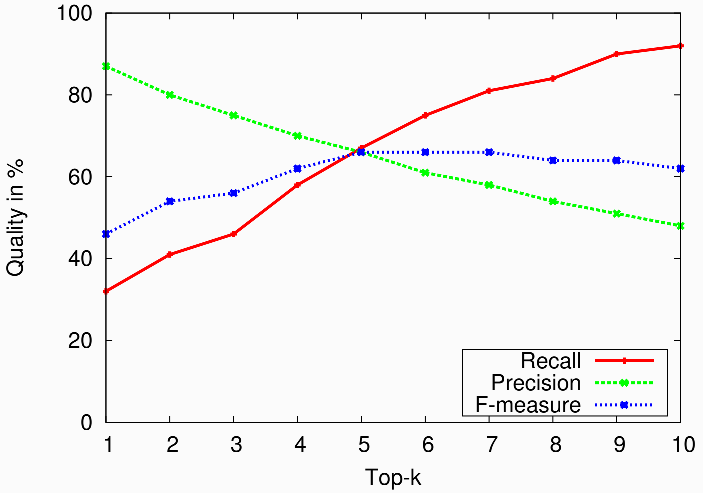
no verification



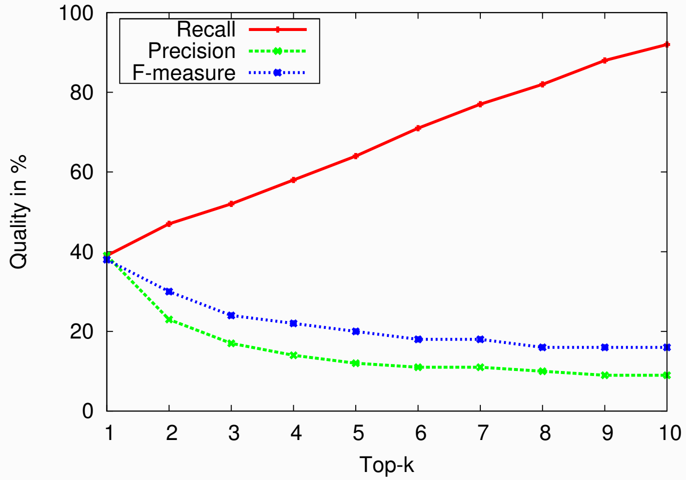
Classification only



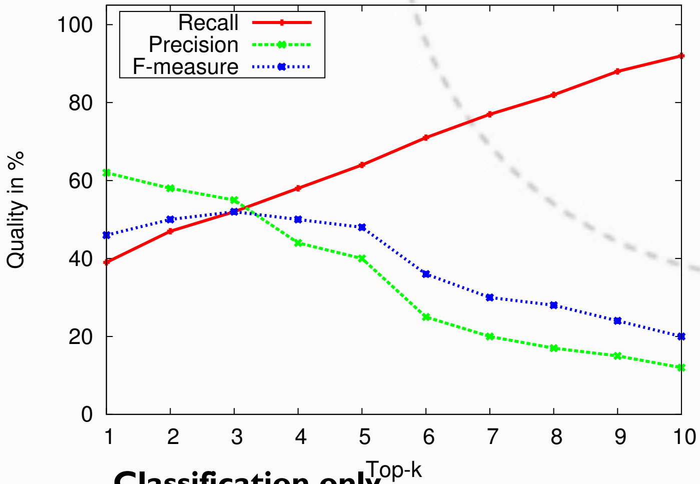
Linking only



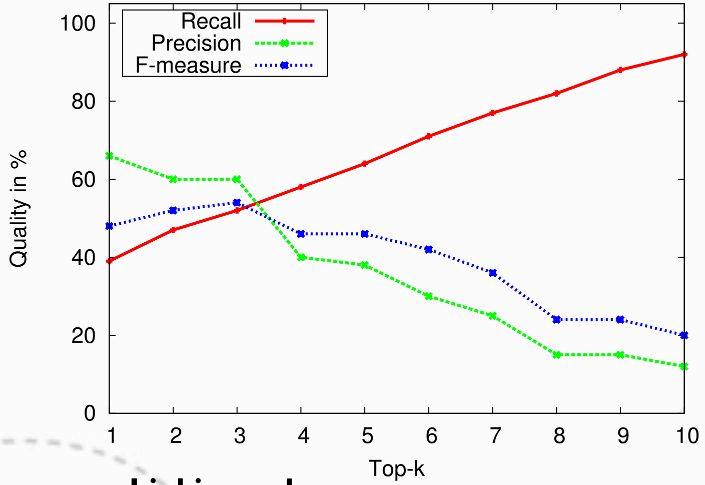
Quality of discovery



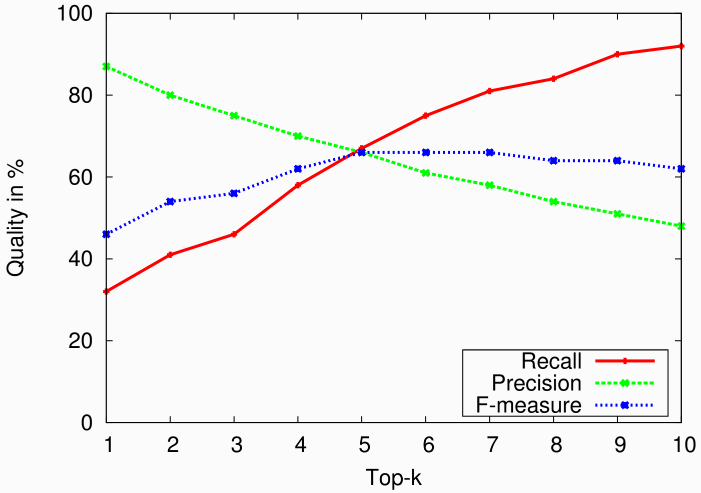
no verification



Classification only



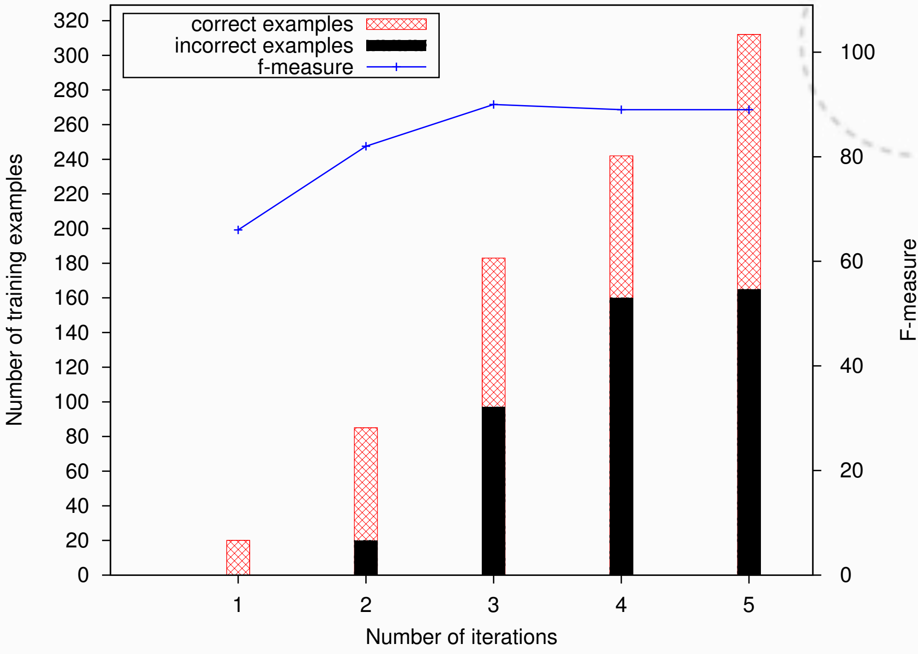
Linking only



with classification and linking

Impact of training data: frequency-based strategy

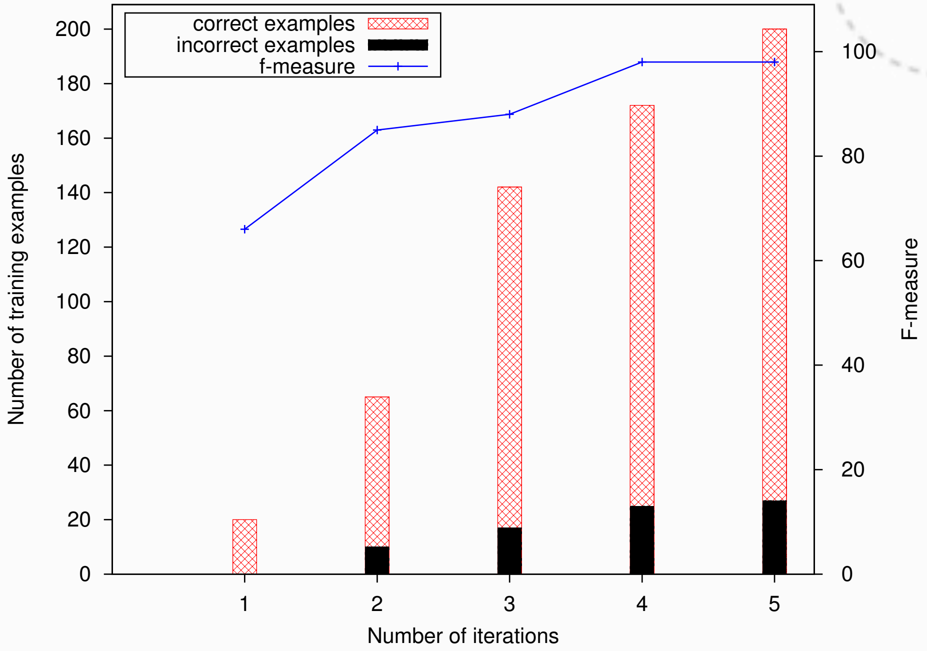
- frequency of a given example discovered in all iterations
- F-measure => best performing classfier at the *i*-th iteration
- classifier m.b. different from one iteration to another (because training data evolves)



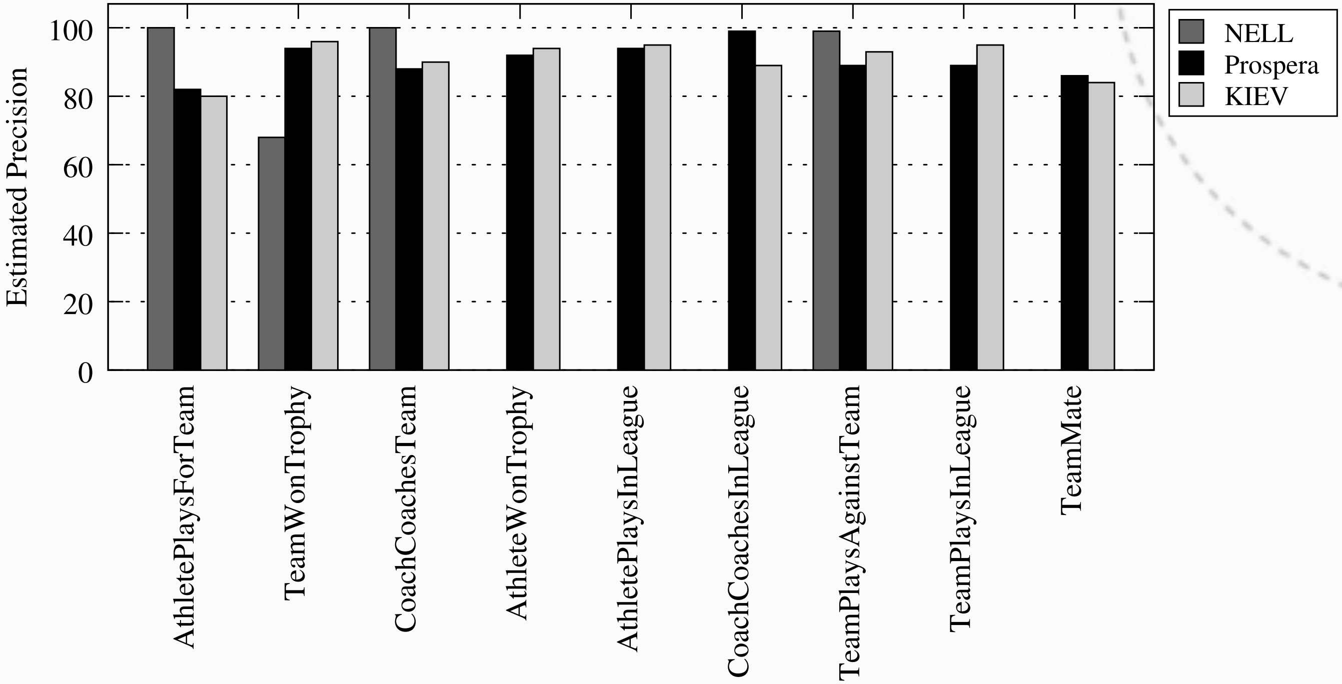
- promotes examples as training data which appear at least 50% of the time in the previous iterations
- the number of examples can grow high
- “stable” F-measure after the 3rd iteration

Impact of training data: linking-based strategy

- enforces harder constraint
- only verified examples are promoted as training examples
- higher F-measure scores
- fewer training data (vs frequency based strategy)
- fewer incorrect examples



Comparative evaluation



- evaluation dataset provided by NELL and Prospera projects and is publicly available
- NELL: 2k facts
- Prospera: 57k facts
- KIEV: 71k facts

Conclusion

- KIEV - populating knowledge bases
- two verification steps
  - classification (to check the type of relationship)
  - linking (to check entities of discovered examples)
- future work
  - experiments from different domains (recently released dataset - ClueWeb2012)
  - study impact of parameters and contradictory cases
  - confidence awareness (exploit provenance info., statistics of patterns)
  - enriching instances with attributes
  - open up the interface and integrate the user feedback (GUI, REST API, and SPARQL endpoint)

Thank you for your attention!

Questions, comments, feedback?

Naimdjon Takhirov
takhirov@idi.ntnu.no