Link Discovery: A Comprehensive Analysis

Nicolai Erbs, Torsten Zesch, Iryna Gurevych
UKP Lab, TU Darmstadt
## Outline

<table>
<thead>
<tr>
<th>Motivation</th>
<th>Link Discovery: A Classification</th>
<th>Evaluation</th>
<th>Conclusions and future work</th>
</tr>
</thead>
</table>

### Motivation

### Link Discovery: A Classification
  - Anchor Discovery
  - Target Discovery
  - Overview

### Evaluation
  - Dataset
  - Anchor Discovery
  - Target Discovery
  - Reducing Links
  - Transfer knowledge from Wikipedia?

### Conclusions and future work
Outline

Motivation

Link Discovery: A Classification
Anchor Discovery
Target Discovery
Overview

Evaluation

Dataset
Anchor Discovery
Target Discovery
Reducing Links
Transfer knowledge from Wikipedia?

Conclusions and future work
Motivation

Links connect web pages

Figure: Wikipedia article, modified from http://en.wikipedia.org/wiki/Chicken_or_the_egg
Motivation

- Links connect web pages
- Quickly navigate from page to page
Motivation

▶ Links connect web pages
▶ Quickly navigate from page to page
▶ Users need motivation to contribute [1]
Motivation

- Links connect web pages
- Quickly navigate from page to page
- Users need motivation to contribute [1]
- Wikipedia: large community of highly motivated users

**Figure:** Wikipedia article, modified from http://en.wikipedia.org/wiki/Chicken_or_the_egg
Motivation

- Links connect web pages
- Quickly navigate from page to page
- Users need motivation to contribute [1]
- Wikipedia: large community of highly motivated users
- Use links for automatic link discovery

**Figure:** Wikipedia article, modified from [http://en.wikipedia.org/wiki/Chicken_or_the_egg](http://en.wikipedia.org/wiki/Chicken_or_the_egg)
Motivation

- What happens if there are no links?

Figure: TWiki article without links
Motivation

▶ What happens if there are no links?
▶ Which comes first, the link or the link discovery?

Figure: TWiki article without links
Motivation

▶ What happens if there are no links?
▶ Which comes first, the link or the link discovery?
▶ Chicken or the egg dilemma

Figure: TWiki article without links

Semantic Web

Semantic Web is a term coined by World Wide Web Consortium (W3C) director Sir Tim Berners-Lee. It describe methods and technologies to allow machines to understand the meaning - or "semantics" - of information on the World Wide Web.

According to the original vision, the availability of machine-readable metadata would enable automated agents and other software to access the Web more intelligently. The agents would be able to perform tasks automatically and locate related information on behalf of the user.
Motivation

- What happens if there are no links?
- Which comes first, the link or the link discovery?
- Chicken or the egg dilemma

- Solution: Text-based link discovery

Figure: TWiki article without links
Outline

Motivation

Link Discovery: A Classification
  Anchor Discovery
  Target Discovery
  Overview

Evaluation
  Dataset
  Anchor Discovery
  Target Discovery
  Reducing Links
  Transfer knowledge from Wikipedia?

Conclusions and future work
Automatic link discovery

1. Select promising link anchors
2. Retrieve best target document

**Figure:** Link discovery approaches split up into a step-by-step representation and classified by the type of knowledge used.
Automatic link discovery
1. Select promising link anchors
2. Retrieve best target document

Prior knowledge
1. Link knowledge
2. Title knowledge
3. Text knowledge

Figure: Link discovery approaches split up into a step-by-step representation and classified by the type of knowledge used.
Anchor Discovery

Approaches

Motivation

Link Discovery: A Classification

Evaluation

Conclusions and future work
Anchor Discovery
Approaches

Link Discovery: A Classification Evaluation

Conclusions and future work

Motivation

Anchor Discovery

Approaches

Link-based

- Major target link score

Formally:

\[ as(a) = \max_d \frac{l(a, d)}{|D_a|} \]  

(1)

\( p \): phrase

\( D \): set of all documents

\( l(a, d) \): # of links from \( a \) to \( d \in D \)

\( D_a \): documents containing \( a \)
Anchor Discovery
Approaches

Link-based
- Major target link score

Formally:
\[ as(a) = \max_d \frac{l(a, d)}{|D_a|} \]  

Title-based
- List of all titles
- Titles are anchors

\begin{itemize}
  \item Refet Bele
  \item Refetoff syndrome
  \item Refeudalisation
  \item Refeudalization
  \item Reffannes
  \item Reffroy
\end{itemize}

$p$: phrase
$D$: set of all documents
$l(a, d)$: # of links from $a$ to $d \in D$
$D_a$: documents containing $a$
Anchor Discovery Approaches

**Link-based**
- Major target link score

Formally:
\[ as(a) = \max_d \frac{l(a, d)}{|D_a|} \]  

\[ p: \text{phrase} \]
\[ D: \text{set of all documents} \]
\[ l(a, d): \# \text{of links from } a \text{ to } d \in D \]
\[ D_a: \text{documents containing } a \]

**Title-based**
- List of all titles
- Titles are anchors

**Text-based**
- Document text only
- Anchor selection
  - Tokens
  - N-grams
  - Noun phrases
- Anchor ranking
  - Cooccurrence graph [2]
  - tf.idf [3]
## Target Discovery

### Approaches

<table>
<thead>
<tr>
<th>Motivation</th>
<th>Link Discovery: A Classification</th>
<th>Evaluation</th>
<th>Conclusions and future work</th>
</tr>
</thead>
</table>

- **Link-based**
  - Most frequent target
  - Formally:
    \[
    t_s(a, d_t) = \sum d_l(a, d_t) \tag{2}
    \]

- **Text-based**
  - Search engine
  - Lucene
  - Terrier

- Following standard IR techniques
  - [http://lucene.apache.org](http://lucene.apache.org)
  - [http://www.terrier.org](http://www.terrier.org)
Target Discovery Approaches

Link Discovery: A Classification Evaluation

Conclusions and future work

Target Discovery

Link-based

▶ Most frequent target

Formally:

\[
ts(a, d_t) = \frac{l(a, d_t)}{\sum_d l(a, d)}
\]  (2)
Target Discovery Approaches

Link-based
- Most frequent target

Formally:
\[ ts(a, d_t) = \frac{l(a, d_t)}{\sum_d l(a, d)} \quad (2) \]

Text-based
- Search engine
  - Lucene\(^a\)
  - Terrier\(^b\)

Following standard IR techniques

\(^a\)http://lucene.apache.org
\(^b\)http://www.terrier.org
**Overview of Approaches**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Knowledge Type</th>
<th>Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICLM</td>
<td>Link</td>
<td>Link anchors</td>
</tr>
<tr>
<td>GPNM</td>
<td>Title and link</td>
<td>Anchor strength</td>
</tr>
<tr>
<td>Text-based</td>
<td>Text</td>
<td>Existing targets</td>
</tr>
</tbody>
</table>

**Knowledge Steps**
- **ICLM system**: LinkText Anchor strength
- **GPNM system**: LinkText Length
- **Text-based system**: Full text search score

**Figure**: Overview of link discovery approaches and the type of knowledge used.
Outline

Motivation

Link Discovery: A Classification
  Anchor Discovery
  Target Discovery
Overview

Evaluation
  Dataset
  Anchor Discovery
  Target Discovery
  Reducing Links
  Transfer knowledge from Wikipedia?

Conclusions and future work
Link Discovery Evaluation

Dataset

- Wikipedia snapshot from October 8, 2008
- Used in the INEX 2009 Link-the-Wiki-Track [6].
- 2,666,190 articles with more than 135 Million links
- Every 1000th article set aside for testing
- Existing links are used as gold standard
Anchor Discovery Evaluation

- Overall precision rather low
Anchor Discovery Evaluation

- Overall precision rather low
- Few links (1% linking ratio)
  - Link-based > text-based
  - Title-based > text-based
Anchor Discovery Evaluation

- Overall precision rather low
- Few links (1% linking ratio)
  - Link-based > text-based
  - Title-based > text-based
- Many links (6% linking ratio)
  - Text-based ≈ link-based
  - Text-based > title-based
Target Discovery Evaluation

- Relaxed version of accuracy
  - 10 target suggestions
  - Correct if one of them matches
  - Similar to users’ view

![Graph showing accuracy comparison between Link-based and Text-based approaches]
Target Discovery Evaluation

- Relaxed version of accuracy
  - 10 target suggestions
  - Correct if one of them matches
  - Similar to users’ view
- Link-based approach performs better than text-based

Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link-based</td>
<td>0.8</td>
</tr>
<tr>
<td>Text-based</td>
<td>0.6</td>
</tr>
</tbody>
</table>
**Target Discovery Evaluation**

- Relaxed version of accuracy
  - 10 target suggestions
  - Correct if one of them matches
  - Similar to users’ view
- Link-based approach performs better than text-based
- Accuracy stays below 0.9 even for 1,000 target suggestions
But, what if there are no links?
Anchor Discovery Evaluation: Reducing Links

- Slowly add links from corpus

**Figure**: Precision of link based anchor discovery depending on the available training data at 6% threshold
Anchor Discovery Evaluation: Reducing Links

- Slowly add links from corpus
- Title-based and text-based approaches are not influenced

**Figure:** Precision of link based anchor discovery depending on the available training data at 6% threshold
Anchor Discovery Evaluation: Reducing Links

- Slowly add links from corpus
- Title-based and text-based approaches are not influenced
- Link-based reaches text-based approach at \( \approx 65 \text{ Million links} \)

**Figure:** Precision of link based anchor discovery depending on the available training data at 6% threshold
Target Discovery Evaluation
Reducing Links

- Slowly add links from corpus

![Graph](image_url)

**Figure:** Accuracy of target discovery depending on the available training data. (Result set size = 5)
Target Discovery Evaluation
Reducing Links

- Slowly add links from corpus
- Text-based approach is not influenced

![Figure](image.png)

**Figure**: Accuracy of target discovery depending on the available training data. (Result set size = 5)
Target Discovery Evaluation
Reducing Links

- Slowly add links from corpus
- Text-based approach is not influenced
- Link-based reaches text-based approach at ≈7 Million links

**Figure:** Accuracy of target discovery depending on the available training data. (Result set size = 5)
<table>
<thead>
<tr>
<th>Motivation</th>
<th>Link Discovery: A Classification</th>
<th>Evaluation</th>
<th>Conclusions and future work</th>
</tr>
</thead>
</table>

**Why not transfer knowledge from Wikipedia?**

- Anchor discovery
  - Using Wikipedia articles may not capture domain-specific anchors
  - Wikipedia does not contain an article for each university professor
  - Good anchor at specific university document collection

- Target discovery
  - Targets can be too specific or general
  - Inside Wikipedia, Java 5 links to Java
  - Should link to Java 5 in more specific collections
Anchor discovery

- Using Wikipedia articles may not capture domain-specific anchors
  - Wikipedia does not contain an article for each university professor
  - Good anchor at specific university document collection
- Product names are only sometimes not worth linking
Why not transfer knowledge from Wikipedia?

Anchor discovery
- Using Wikipedia articles may not capture domain-specific anchors
  - Wikipedia does not contain an article for each university professor
  - Good anchor at specific university document collection
- Product names are only sometimes not worth linking

Target Discovery
- Targets can be too specific or general
  - Inside Wikipedia Java 5 links to Java
  - Should link to Java 5 in more specific collections
Outline

Motivation

Link Discovery: A Classification
   Anchor Discovery
   Target Discovery
   Overview

Evaluation
   Dataset
   Anchor Discovery
   Target Discovery
   Reducing Links
   Transfer knowledge from Wikipedia?

Conclusions and future work
Conclusions and future work

- Link-based approach performs best for Wikipedia
Conclusions and future work

- Link-based approach performs best for Wikipedia
- Link-based and title-based approaches cannot easily be transferred to other document collections
Conclusions and future work

- Link-based approach performs best for Wikipedia
- Link-based and title-based approaches cannot easily be transferred to other document collections
- Link-based approach does not work if few links are available
Conclusions and future work

- Link-based approach performs best for Wikipedia
- Link-based and title-based approaches cannot easily be transferred to other document collections
- Link-based approach does not work if few links are available
- Text-based approaches can be used for reliable link discovery in arbitrary document collections
Conclusions and future work

- Link-based approach performs best for Wikipedia
- Link-based and title-based approaches cannot easily be transferred to other document collections
- Link-based approach does not work if few links are available

- Text-based approaches can be used for reliable link discovery in arbitrary document collections
- Combine all approaches for best link discovery


<table>
<thead>
<tr>
<th>Motivation</th>
<th>Link Discovery: A Classification</th>
<th>Evaluation</th>
<th>Conclusions and future work</th>
</tr>
</thead>
</table>


Thank you for your attention