Link Discovery: A Comprehensive Analysis

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

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

Outline

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Motivation

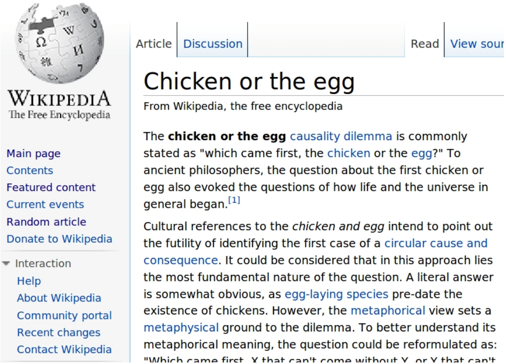
Link Discovery: A Classification
- Anchor Discovery
- Target Discovery
- Overview

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

Conclusions and future work

Motivation

- Links connect web pages



Chicken or the egg

From Wikipedia, the free encyclopedia

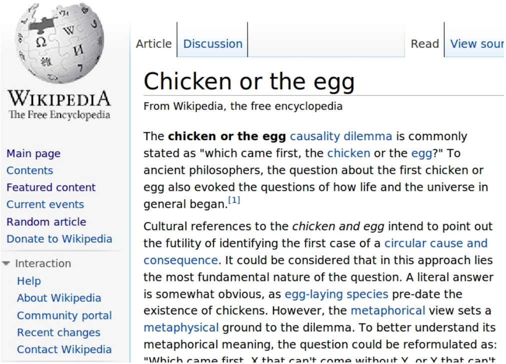
The **chicken or the egg** causality dilemma is commonly stated as "which came first, the chicken or the egg?" To ancient philosophers, the question about the first chicken or egg also evoked the questions of how life and the universe in general began.[1]

Cultural references to the *chicken and egg* intend to point out the futility of identifying the first case of a circular cause and consequence. It could be considered that in this approach lies the most fundamental nature of the question. A literal answer is somewhat obvious, as egg-laying species pre-date the existence of chickens. However, the metaphorical view sets a metaphysical ground to the dilemma. To better understand its metaphorical meaning, the question could be reformulated as: "Which came first, X that can't come without Y, or Y that can't

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



Chicken or the egg

From Wikipedia, the free encyclopedia

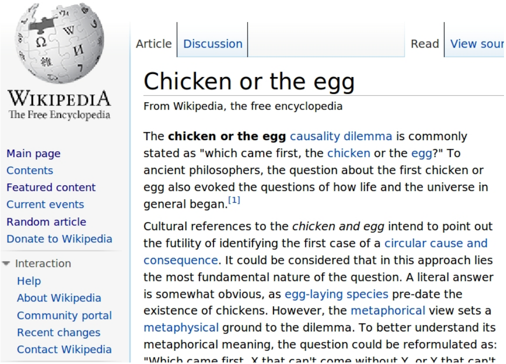
The **chicken or the egg** causality dilemma is commonly stated as "which came first, the chicken or the egg?" To ancient philosophers, the question about the first chicken or egg also evoked the questions of how life and the universe in general began.[1]

Cultural references to the *chicken and egg* intend to point out the futility of identifying the first case of a circular cause and consequence. It could be considered that in this approach lies the most fundamental nature of the question. A literal answer is somewhat obvious, as egg-laying species pre-date the existence of chickens. However, the metaphorical view sets a metaphysical ground to the dilemma. To better understand its metaphorical meaning, the question could be reformulated as: "Which came first, X that can't come without Y, or Y that can't

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]



Chicken or the egg

From Wikipedia, the free encyclopedia

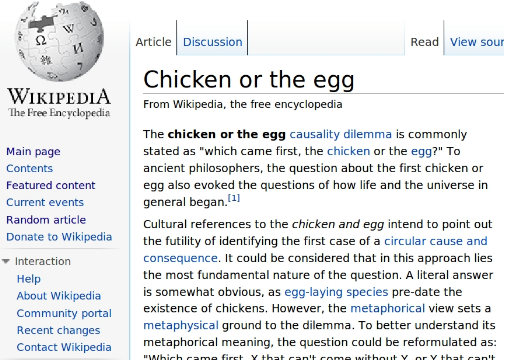
The **chicken or the egg** causality dilemma is commonly stated as "which came first, the chicken or the egg?" To ancient philosophers, the question about the first chicken or egg also evoked the questions of how life and the universe in general began.[1]

Cultural references to the *chicken and egg* intend to point out the futility of identifying the first case of a circular cause and consequence. It could be considered that in this approach lies the most fundamental nature of the question. A literal answer is somewhat obvious, as egg-laying species pre-date the existence of chickens. However, the metaphorical view sets a metaphysical ground to the dilemma. To better understand its metaphorical meaning, the question could be reformulated as: "Which came first, X that can't come without Y, or Y that can't

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



Chicken or the egg

From Wikipedia, the free encyclopedia

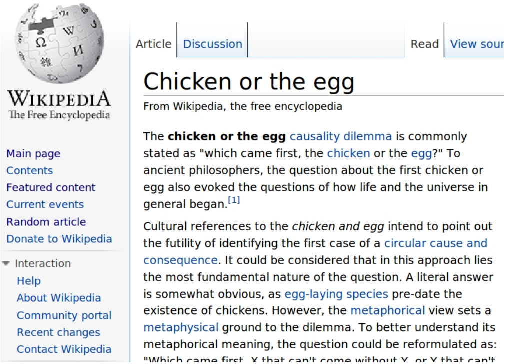
The **chicken or the egg** causality dilemma is commonly stated as "which came first, the chicken or the egg?" To ancient philosophers, the question about the first chicken or egg also evoked the questions of how life and the universe in general began.[1]

Cultural references to the *chicken and egg* intend to point out the futility of identifying the first case of a circular cause and consequence. It could be considered that in this approach lies the most fundamental nature of the question. A literal answer is somewhat obvious, as egg-laying species pre-date the existence of chickens. However, the metaphorical view sets a metaphysical ground to the dilemma. To better understand its metaphorical meaning, the question could be reformulated as: "Which came first, X that can't come without Y, or Y that can't

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



Chicken or the egg

From Wikipedia, the free encyclopedia

The **chicken or the egg** causality dilemma is commonly stated as "which came first, the chicken or the egg?" To ancient philosophers, the question about the first chicken or egg also evoked the questions of how life and the universe in general began.[1]

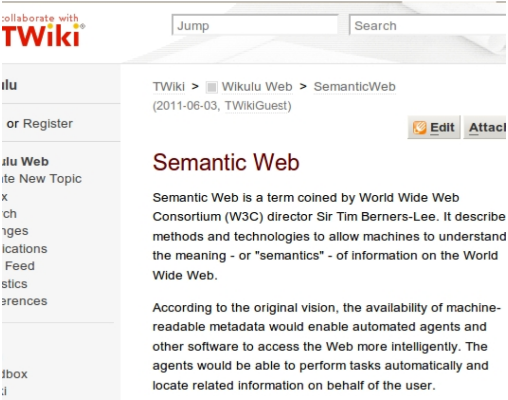
Cultural references to the *chicken and egg* intend to point out the futility of identifying the first case of a circular cause and consequence. It could be considered that in this approach lies the most fundamental nature of the question. A literal answer is somewhat obvious, as egg-laying species pre-date the existence of chickens. However, the metaphorical view sets a metaphysical ground to the dilemma. To better understand its metaphorical meaning, the question could be reformulated as: "Which came first, X that can't come without Y, or Y that can't

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

Motivation

TECHNISCHE
UNIVERSITÄT
DARMSTADT

- What happens if there are no links?



Semantic Web

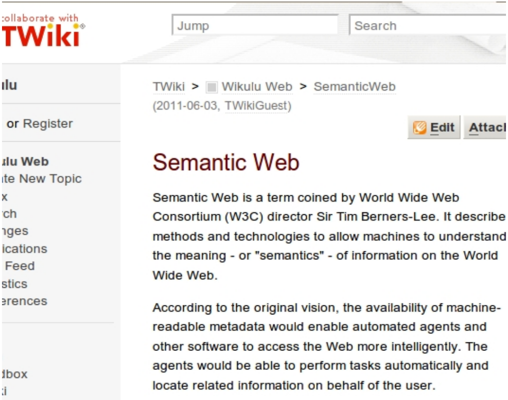
Semantic Web is a term coined by World Wide Web Consortium (W3C) director Sir Tim Berners-Lee. It describe methods and technologies to allow machines to understand the meaning - or "semantics" - of information on the World Wide Web.

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.

Figure: TWiki article without links

Motivation

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



Semantic Web

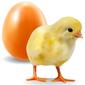
Semantic Web is a term coined by World Wide Web Consortium (W3C) director Sir Tim Berners-Lee. It describe methods and technologies to allow machines to understand the meaning - or "semantics" - of information on the World Wide Web.

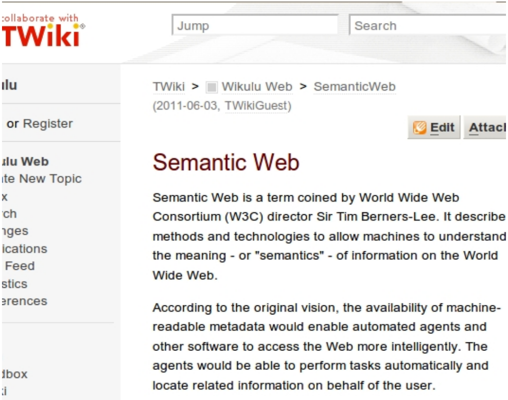
According to the original vision, the availability of machine-readable metadata would enable automated agents and other software to access the Web more intelligently. The agents would be able to perform tasks automatically and locate related information on behalf of the user.

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*





TWiki > Wikulu Web > SemanticWeb
(2011-06-03, TWikiGuest)

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.

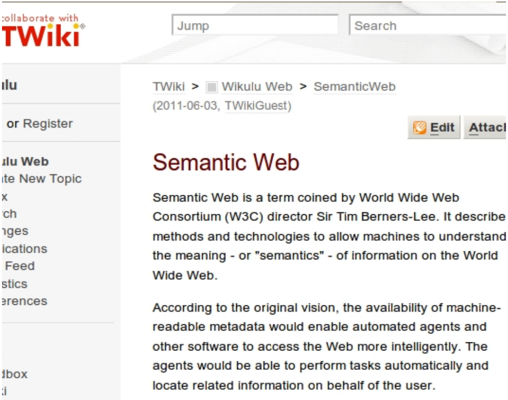
According to the original vision, the availability of machine-readable metadata would enable automated agents and other software to access the Web more intelligently. The agents would be able to perform tasks automatically and locate related information on behalf of the user.

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*

- Solution:
  Text-based link discovery



TWiki > Wikulu Web > SemanticWeb
(2011-06-03, TWikiGuest)

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.

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

Link Discovery
A Classification

- Automatic link discovery
  1. Select promising link anchors
  2. Retrieve best target document

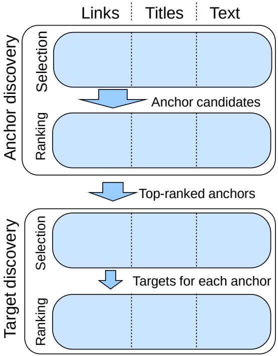


Figure: Link discovery approaches split up into a step-by-step representation and classified by the type of knowledge used.

Link Discovery
A Classification

- Automatic link discovery
  1. Select promising link anchors
  2. Retrieve best target document
- Prior knowledge
  1. Link knowledge
  2. Title knowledge
  3. Text knowledge

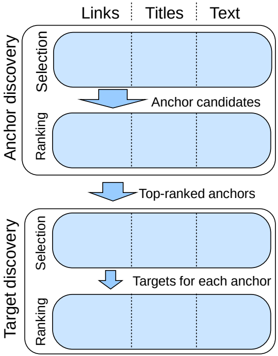


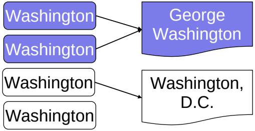
Figure: Link discovery approaches split up into a step-by-step representation and classified by the type of knowledge used.

Anchor Discovery Approaches

Anchor Discovery Approaches

Link-based

- Major target link score



Formally:

$$as(a) = max_d \frac{l(a,d)}{|D_a|} \qquad (1)$$

p: phrase

D: set of all documents

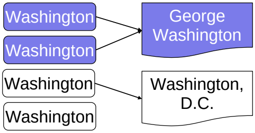
$l(a, d)$: # of links from a to $d \in D$

D_a: documents containing a

Anchor Discovery Approaches

Link-based

- Major target link score



Formally:

$$as(a) = max_d \frac{l(a,d)}{|D_a|} \qquad (1)$$

p: phrase

D: set of all documents

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

D_a: documents containing a

Title-based

- List of all titles
- Titles are anchors

Refet Bele

Refetoff syndrome

Refeudalisation

Refeudalization

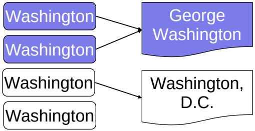
Reffannes

Reffroy

Anchor Discovery Approaches

Link-based

- Major target link score



Formally:

$$as(a) = max_d \frac{l(a,d)}{|D_a|} \qquad (1)$$

p: phrase

D: set of all documents

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

D_a: documents containing a

Title-based

- List of all titles
- Titles are anchors

Refet Bele

Refetoff syndrome

Refeudalisation

Refeudalization

Reffannes

Reffroy

Text-based

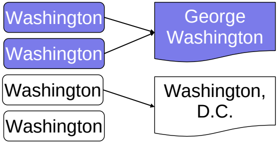
- Document text only
- Anchor selection
  - Tokens
  - N-grams
  - Noun phrases
- Anchor ranking
  - Cooccurrence graph [2]
  - tf.idf [3]

Target Discovery Approaches

Target Discovery Approaches

Link-based

- Most frequent target



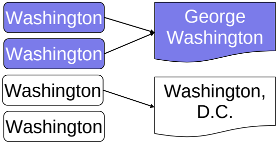
Formally:

$$ts(a, d_t) = \frac{l(a, d_t)}{\Sigma_d \, l(a, d)} \qquad (2)$$

Target Discovery Approaches

Link-based

- Most frequent target



Formally:

$$ts(a, d_t) = \frac{l(a, d_t)}{\sum_d l(a, d)} \qquad (2)$$

Text-based

- Search engine
  - Lucene[a]
  - Terrier[b]

Following standard IR techniques

[a] `http://lucene.apache.org`

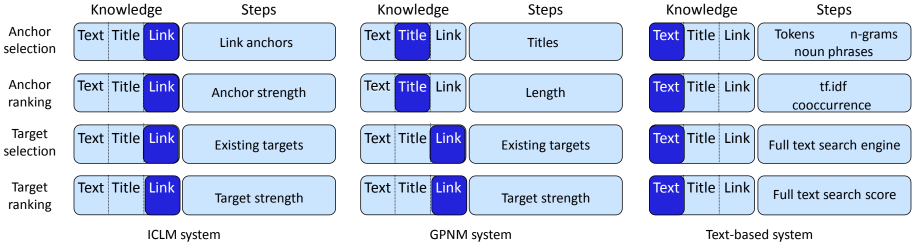
[b] `http://www.terrier.org`

Overview of Approaches

ICLM Relies on link knowledge [4]

GPNM Combines title and link knowledge [5]

Text-based Uses only the document text



	ICLM system		GPNM system		Text-based system	
	Knowledge	Steps	Knowledge	Steps	Knowledge	Steps
Anchor selection	Text Title **Link**	Link anchors	Text **Title** Link	Titles	**Text** Title Link	Tokens n-grams noun phrases
Anchor ranking	Text Title **Link**	Anchor strength	Text **Title** Link	Length	**Text** Title Link	tf.idf cooccurrence
Target selection	Text Title **Link**	Existing targets	Text Title **Link**	Existing targets	**Text** Title Link	Full text search engine
Target ranking	Text Title **Link**	Target strength	Text Title **Link**	Target strength	**Text** Title Link	Full text search score

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

Outline

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Motivation

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

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

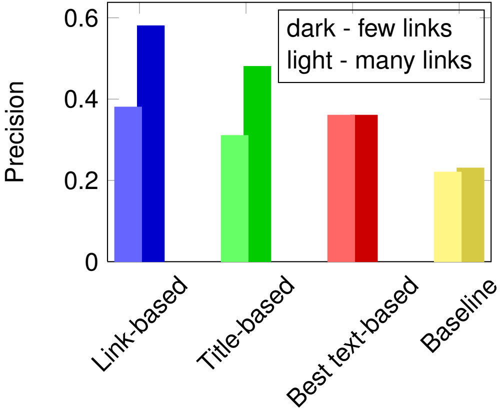
Conclusions and future work

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

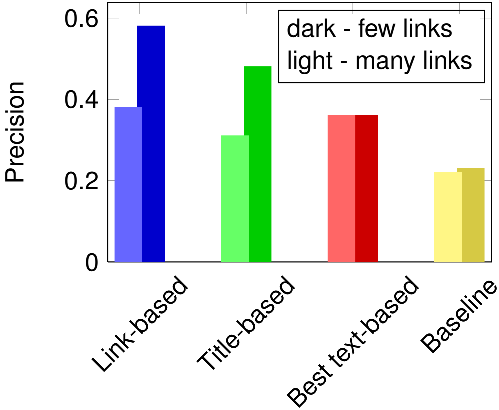
Anchor Discovery Evaluation

- Overall precision rather low



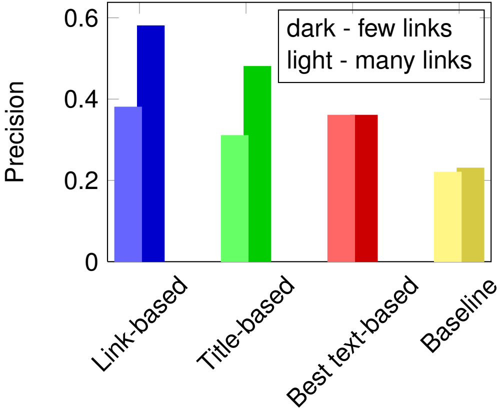
Anchor Discovery Evaluation

- Overall precision rather low
- few links (1% linking ratio)
  Link-based > text-based
  Title-based > text-based



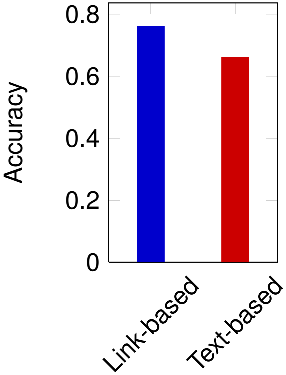
Anchor Discovery Evaluation

- Overall precision rather low
- few links (1% linking ratio)
  Link-based $>$ text-based
  Title-based $>$ text-based
- many links (6% linking ratio)
  Text-based $\approx$ link-based
  Text-based $>$ title-based



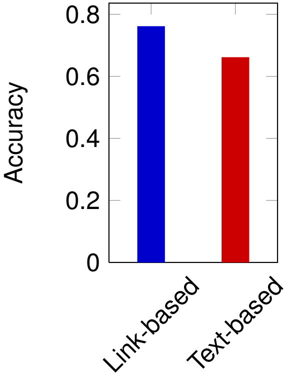
Target Discovery Evaluation

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



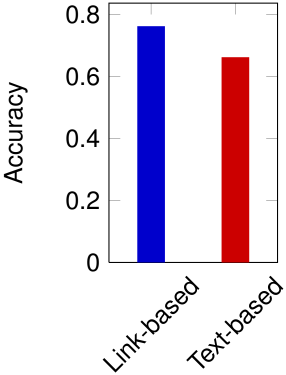
Target Discovery Evaluation

- 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

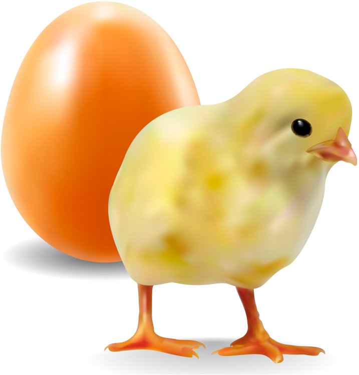


Target Discovery Evaluation

- 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

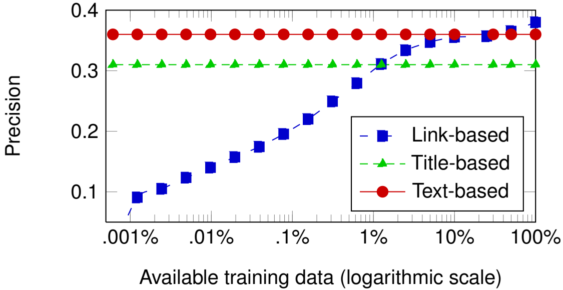


Figure: Precision of link based anchor discovery depending on the available training data at 6% threshold

Anchor Discovery Evaluation: Reducing Links

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

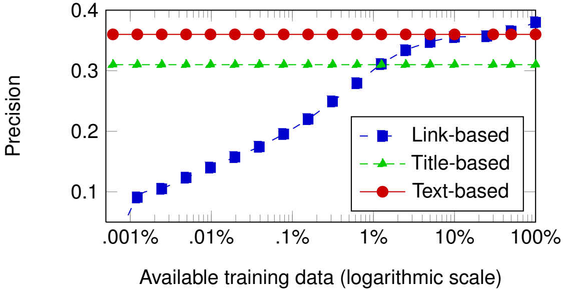


Figure: Precision of link based anchor discovery depending on the available training data at 6% threshold

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 ≈65 Million links

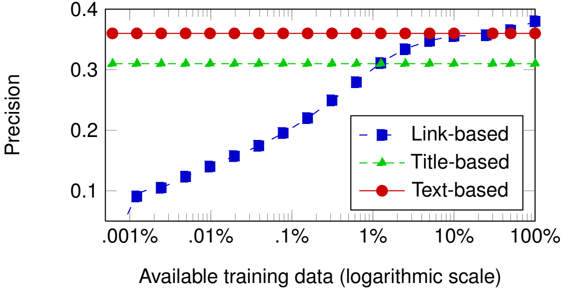


Figure: Precision of link based anchor discovery depending on the available training data at 6% threshold

Target Discovery Evaluation
Reducing Links

- Slowly add links from corpus

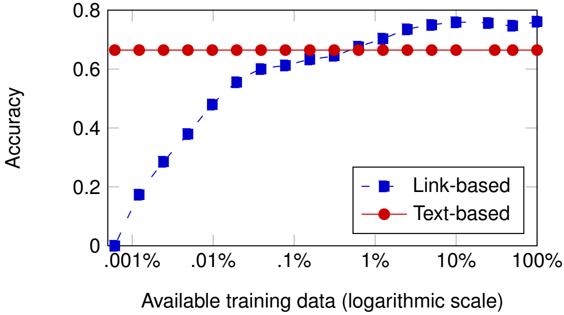


Figure: Accuracy of target discovery depending on the available training data. (Result set size = 5)

Target Discovery Evaluation
Reducing Links

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

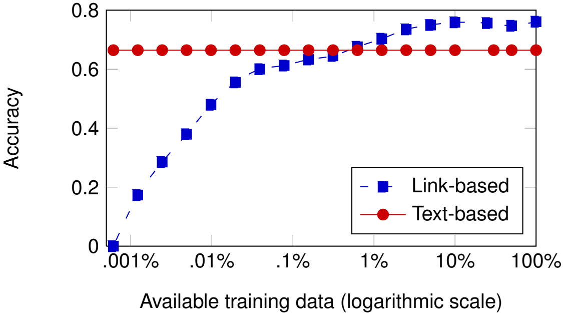


Figure: Accuracy of target discovery depending on the available training data. (Result set size = 5)

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

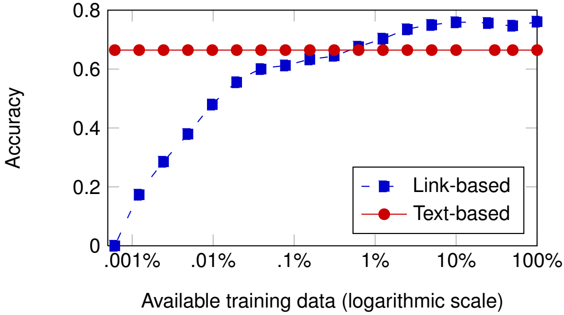


Figure: Accuracy of target discovery depending on the available training data. (Result set size = 5)

Why not transfer knowledge from Wikipedia?

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

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

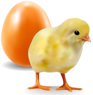
- Link-based approach performs best for Wikipedia

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

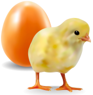
Conclusions and future work

- 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

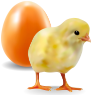
- Link-based approach performs best for Wikipedia
- Link-based and title-based approaches cannot easily be transferred to other document collections
- Link-based approach does not work if few links are available



- 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

References I

- A. Majchrzak, C. Wagner, and D. Yates, “Corporate Wiki Users: Results of a Survey,” in *WikiSym ’06: Proceedings of the 2006 international symposium on Wikis*, Odense, Denmark, 2006, pp. 99–104. [Online]. Available: http://doi.acm.org/10.1145/1149453.1149472

- R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Text,” in *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 404–411. [Online]. Available: http://www.aclweb.org/anthology/W04-3252

- G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval.* McGraw Hill, 1983.

- K. Y. Itakura and C. L. A. Clarke, “University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks,” in *INEX 2007 Workshop Preproceedings*, vol. 4862, 2007, pp. 417–425. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-85902-4_35

References II

- S. Geva, “GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia,” in *Preproceedings of the INEX Workshop*, 2007, pp. 404–416.
- D. W. C. Huang, S. Geva, and A. Trotman, “Overview of the INEX 2009 Link the Wiki Track,” in *INEX*, ser. Lecture Notes in Computer Science, vol. 6203, 2009, pp. 312–323. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-14556-8

Thank you for your attention

KTS
KLAUS TSCHIRA STIFTUNG GEMEINNÜTZIGE GMBH

e-learning center of research excellence

Bundesministerium für Wirtschaft und Technologie

VolkswagenStiftung

Deutsche Forschungsgemeinschaft DFG

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG

Ubiquitous Knowledge Processing Lab

`http://www.ukp.tu-darmstadt.de`