A Cross-Lingual Dictionary for English Wikipedia Concepts

Valentin I. Spitkovsky

with **Angel X. Chang**

Stanford University / Google Inc.

From Words to Concepts and Back:

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

Yet in each word some concept there must be...

— from Goethe's *Faust*

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

Yet in each word some concept there must be...

— from Goethe's *Faust*

Example:

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

Yet in each word some concept there must be...

— from Goethe's *Faust*

Example:

- **word sense disambiguation**

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

Yet in each word some concept there must be...

— from Goethe's *Faust*

Example:

- **word sense disambiguation**

football

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

Yet in each word some concept there must be...

— from Goethe's *Faust*

Example:

- **word sense disambiguation**

football

From Words to Concepts and Back:

Dictionaries for Linking Text, Entities and Ideas

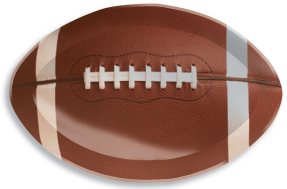
Yet in each word some concept there must be...

— from Goethe's *Faust*

Example:

- **word sense disambiguation**

football



Problem Space:

Problem Space:

- **words:**

Problem Space:

- **words: raw, unstructured natural language representation**

Problem Space:

- **words: raw, unstructured natural language representation**
  - **low-level (high-dimensional)**

Problem Space:

- **words: raw, unstructured natural language representation**
  - **low-level (high-dimensional)**
- **concepts:**

Problem Space:

- **words: raw, unstructured natural language representation**
  - **low-level (high-dimensional)**
- **concepts: concrete, structured organization of knowledge**

Problem Space:

- **words: raw, unstructured natural language representation**
  - **low-level (high-dimensional)**
- **concepts: concrete, structured organization of knowledge**
  - **Wikipedia articles**

Problem Space:

- **words: raw, unstructured natural language representation**
  - **low-level (high-dimensional)**
- **concepts: concrete, structured organization of knowledge**
  - **Wikipedia articles, as in explicit semantic analysis (ESA)**
    (Gabrilovich and Markovitch, 2007)

Problem Space:

- **words: raw, unstructured natural language representation**
  - **low-level (high-dimensional)**
- **concepts: concrete, structured organization of knowledge**
  - **Wikipedia articles, as in explicit semantic analysis (ESA)**
    (Gabrilovich and Markovitch, 2007)
- **or coarse categories**

Problem Space:

- **words: raw, unstructured natural language representation**
  - **low-level (high-dimensional)**
- **concepts: concrete, structured organization of knowledge**
  - **Wikipedia articles, as in explicit semantic analysis (ESA)**
    (Gabrilovich and Markovitch, 2007)
- **or coarse categories**
  - **high-level (low-dimensional) representation**

Problem Space:

- **words: raw, unstructured natural language representation**
  - **low-level (high-dimensional)**
- **concepts: concrete, structured organization of knowledge**
  - **Wikipedia articles, as in explicit semantic analysis (ESA)**
    (Gabrilovich and Markovitch, 2007)
- **or coarse categories**
  - **high-level (low-dimensional) representation**
  - **e.g., aggregation via Wikipedia's hierarchical structure**

Connection:

Connection:

Leech's main academic interests are: English grammar; ...

Corpus-based natural language processing by computer

Connection:

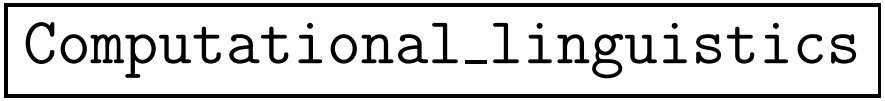
Leech's main academic interests are: English grammar; ...

Corpus-based natural language processing by computer

Connection:

Leech's main academic interests are: English grammar; ...

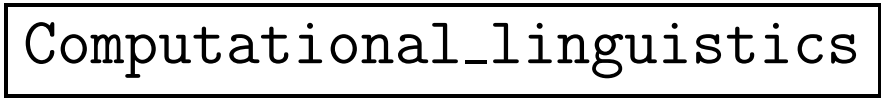
Corpus-based natural language processing by computer



Connection:

Leech's main academic interests are: English grammar; ...

Corpus-based natural language processing by computer

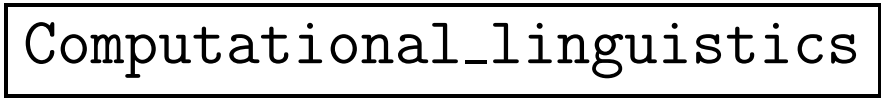


He is also a computational linguist who...

Connection:

Leech's main academic interests are: English grammar; ...

Corpus-based natural language processing by computer



He is also a computational linguist who...

Connection:

Leech's main academic interests are: English grammar; ...

Corpus-based natural language processing by computer

Computational_linguistics

Computerlinguistik

He is also a computational linguist who...

Connection:

Leech's main academic interests are: English grammar; ...

Corpus-based natural language processing by computer

Computational_linguistics

Computerlinguistik

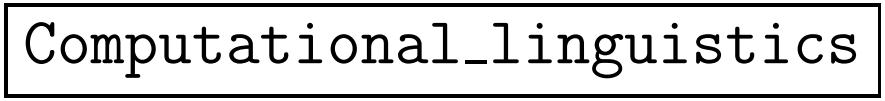
Linguistique informatique

He is also a computational linguist who...

Connection:

Leech's main academic interests are: English grammar; ...

Corpus-based natural language processing by computer



Computerlinguistik

Linguistique informatique

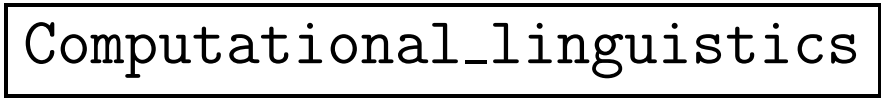
Språkteknologi

He is also a computational linguist who...

Connection:

Leech's main academic interests are: English grammar; ...

Corpus-based natural language processing by computer



Computerlinguistik

Linguistique informatique

Språkteknologi

⋮

He is also a computational linguist who...

Solution:

Solution:

- **anchor-texts are pretty good descriptors of pages**

(Manning, Raghavan and Schütze, 2008; Ch. 21)

Solution:

- anchor-texts are pretty good descriptors of pages

  (Manning, Raghavan and Schütze, 2008; Ch. 21)

- collect all anchor-text from each article's incoming links

Solution:

- anchor-texts are pretty good descriptors of pages

(Manning, Raghavan and Schütze, 2008; Ch. 21)

- collect all anchor-text from each article's incoming links

$$\{(\text{concept}, \text{words}) \mapsto \text{count}\}$$

Solution:

- anchor-texts are pretty good descriptors of pages

(Manning, Raghavan and Schütze, 2008; Ch. 21)

- collect all anchor-text from each article's incoming links

$$\{(\textbf{concept}, \textbf{words}) \mapsto \textbf{count}\}$$

$$\hat{\mathbb{P}}(\textbf{concept} \mid \textbf{words}) = \frac{\textbf{count}(\textbf{concept}, \textbf{words})}{\sum \textbf{count}(*, \textbf{words})}$$

Solution:

- anchor-texts are pretty good descriptors of pages

(Manning, Raghavan and Schütze, 2008; Ch. 21)

- collect all anchor-text from each article's incoming links

$$\{(\textbf{concept}, \textbf{words}) \mapsto \textbf{count}\}$$

$$\hat{\mathbb{P}}(\textbf{concept} \mid \textbf{words}) = \frac{\textbf{count}(\textbf{concept}, \textbf{words})}{\sum \textbf{count}(*, \textbf{words})}$$

$$\hat{\mathbb{P}}(\textbf{words} \mid \textbf{concept}) = \frac{\textbf{count}(\textbf{concept}, \textbf{words})}{\sum \textbf{count}(\textbf{concept}, *)}$$

Types:

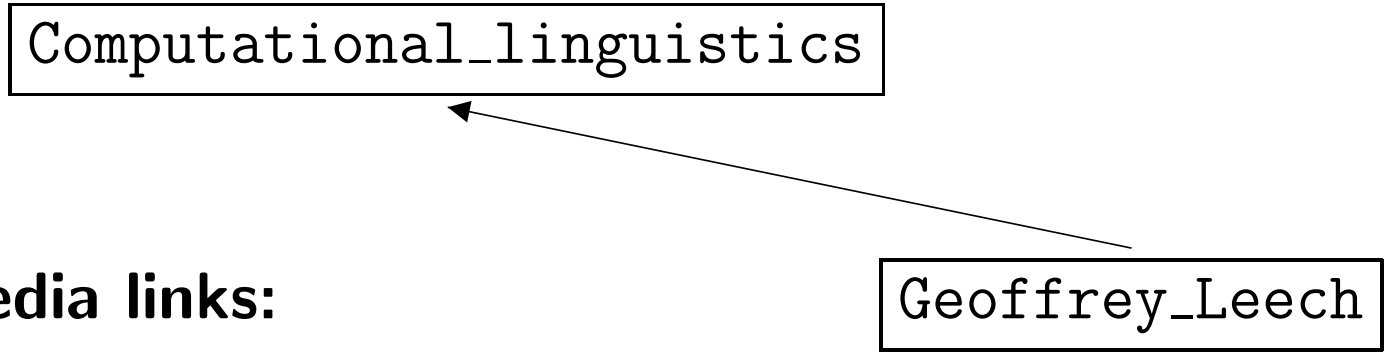
Computational_linguistics

Types:

`Computational_linguistics`

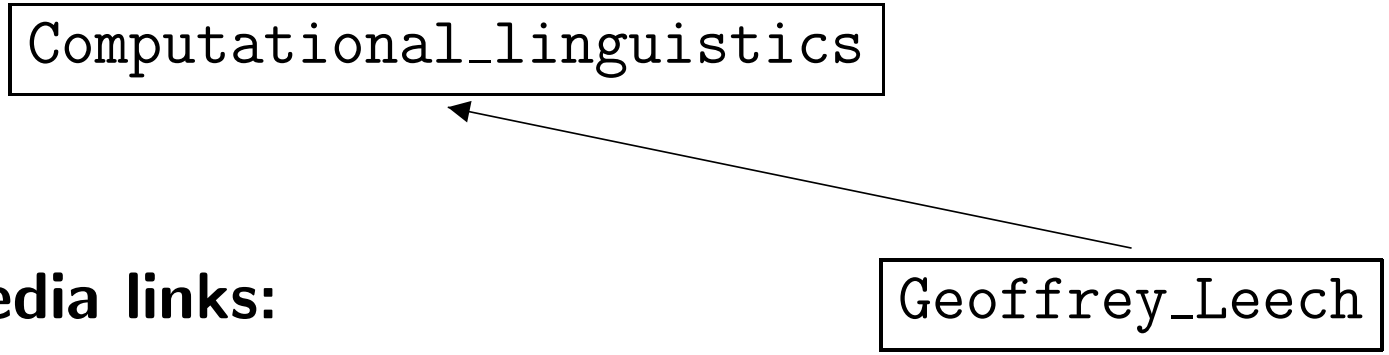
1. **inter-Wikipedia links:**

Types:



1. **inter-Wikipedia links:**

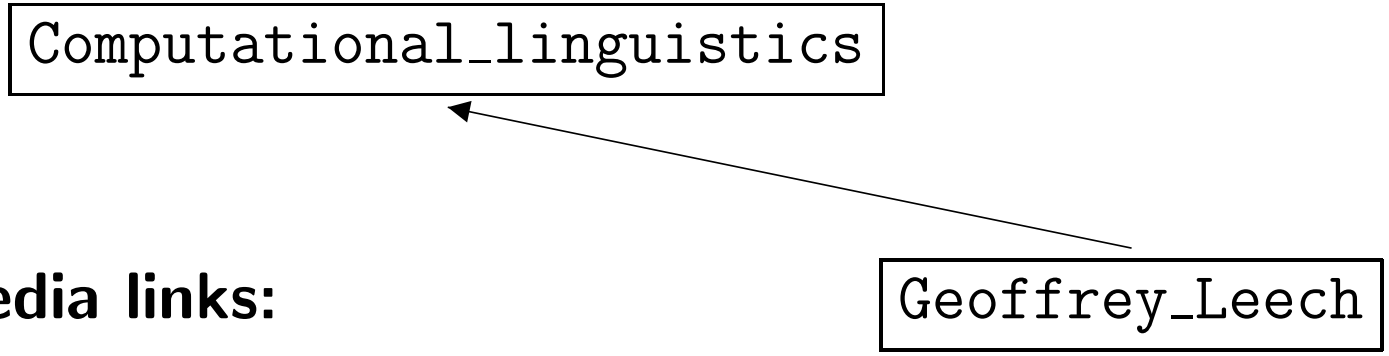
Types:



1. **inter-Wikipedia links:**

   Leech's main academic interests are: English grammar;
   ... Corpus-based natural language processing by computer

Types:

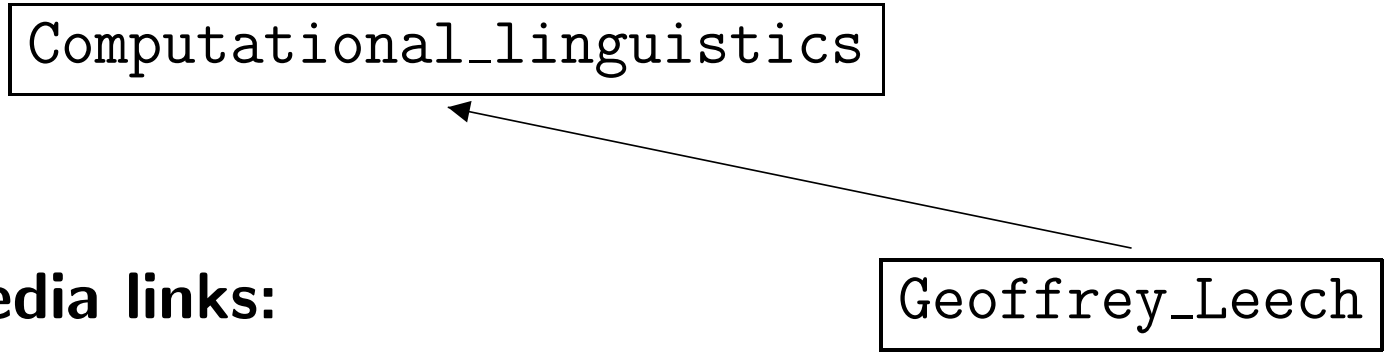


1. **inter-Wikipedia links:**

   Leech's main academic interests are: English grammar; ... Corpus-based natural language processing by computer

2. **external links:**

Types:



1. **inter-Wikipedia links:**

   Leech's main academic interests are: English grammar; ... Corpus-based natural language processing by computer

2. **external links:** www.culinaryanthropologist.org/about.html

Types:

Computational_linguistics

Geoffrey_Leech

1. **inter-Wikipedia links:**

   Leech's main academic interests are: English grammar; ... Corpus-based natural language processing by computer

2. **external links:** www.culinaryanthropologist.org/about.html

   Matt eats very well. He is also a computational linguist who takes time off from the research he usually does for culinary road trips and other adventures.

Cross-lingual Examples:

`Computational_linguistics`

Cross-lingual Examples:

`Computational_linguistics`

3. **anchor-texts of links into parallel Wikipedia pages:**

Cross-lingual Examples:

`Computational_linguistics`

3. **anchor-texts of links into parallel Wikipedia pages:**
   - `de/Computerlinguistik`

Cross-lingual Examples:

`Computational_linguistics`

3. **anchor-texts of links into parallel Wikipedia pages:**
   - `de/Computerlinguistik`
   - `fr/Linguistique_informatique`

Cross-lingual Examples:

`Computational_linguistics`

3. **anchor-texts of links into parallel Wikipedia pages:**
   - `de/Computerlinguistik`
   - `fr/Linguistique_informatique`
   - `sv/Språkteknologi`

Cross-lingual Examples:

`Computational_linguistics`

3. **anchor-texts of links into parallel Wikipedia pages:**
   - `de/Computerlinguistik`
   - `fr/Linguistique_informatique`
   - `sv/Språkteknologi`
4. **... titles and other relevant strings!**

Cross-lingual Examples:

`Computational_linguistics`

3. **anchor-texts of links into parallel Wikipedia pages:**
   - `de/Computerlinguistik`
   - `fr/Linguistique_informatique`
   - `sv/Språkteknologi`
4. **... titles and other relevant strings!** **(these don't count)**

Volume:

Volume:

- **wisdom of one huge crowd!**

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links**

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links** (∼ **half English, half parallel**)

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links** (∼ **half English, half parallel**)
  - **297,073,139 distinct concept-word pairs**

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links** (∼ **half English, half parallel**)
  - **297,073,139 distinct concept-word pairs**
  - **175,100,788 unique strings**

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links** (∼ **half English, half parallel**)
  - **297,073,139 distinct concept-word pairs**
  - **175,100,788 unique strings**
  - **7,560,141 concepts**

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links** (∼ **half English, half parallel**)
  - **297,073,139 distinct concept-word pairs**
  - **175,100,788 unique strings**
  - **7,560,141 concepts**

- **includes "red" links to non-existent pages...**

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links** (∼ **half English, half parallel**)
  - **297,073,139 distinct concept-word pairs**
  - **175,100,788 unique strings**
  - **7,560,141 concepts**

- **includes "red" links to non-existent pages...**
  - **canonicalize everything** (**especially redirects**)

Volume:

- wisdom of one huge crowd!
  - 3,152,091,432 individual links (∼ half English, half parallel)
  - 297,073,139 distinct concept-word pairs
  - 175,100,788 unique strings
  - 7,560,141 concepts
- includes "red" links to non-existent pages...
  - canonicalize everything (especially redirects)
- Wikipedia's coverage is extensive (and growing)

Volume:

- wisdom of one huge crowd!
  - 3,152,091,432 individual links (∼ half English, half parallel)
  - 297,073,139 distinct concept-word pairs
  - 175,100,788 unique strings
  - 7,560,141 concepts
- includes "red" links to non-existent pages...
  - canonicalize everything (especially redirects)
- Wikipedia's coverage is extensive (and growing)
  - extrinsic quantity → quality

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links** (∼ **half English, half parallel**)
  - **297,073,139 distinct concept-word pairs**
  - **175,100,788 unique strings**
  - **7,560,141 concepts**
- **includes "red" links to non-existent pages...**
  - **canonicalize everything** (**especially redirects**)
- **Wikipedia's coverage is extensive** (**and growing**)
  - **extrinsic quantity → quality**
  - ***not* intrinsic quality** (**main differentiator**)

Volume:

- wisdom of one huge crowd!
  - 3,152,091,432 individual links (∼ half English, half parallel)
  - 297,073,139 distinct concept-word pairs
  - 175,100,788 unique strings
  - 7,560,141 concepts
- includes "red" links to non-existent pages...
  - canonicalize everything (especially redirects)
- Wikipedia's coverage is extensive (and growing)
  - extrinsic quantity → quality
  - *not* intrinsic quality (main differentiator)

  — pre-Wikipedia (Koningstein et al., 2003–4)

Volume:

- wisdom of one huge crowd!
  - 3,152,091,432 individual links (∼ half English, half parallel)
  - 297,073,139 distinct concept-word pairs
  - 175,100,788 unique strings
  - 7,560,141 concepts
- includes "red" links to non-existent pages...
  - canonicalize everything (especially redirects)
- Wikipedia's coverage is extensive (and growing)
  - extrinsic quantity → quality
  - *not* intrinsic quality (main differentiator)

    — pre-Wikipedia (Koningstein et al., 2003–4),

    ESA (2007), Wiki-linking (Milne and Witten, 2008), etc.

Volume:

- **wisdom of one huge crowd!**
  - **3,152,091,432 individual links** (∼ **half English, half parallel**)
  - **297,073,139 distinct concept-word pairs**
  - **175,100,788 unique strings**
  - **7,560,141 concepts**

- **includes "red" links to non-existent pages...**
  - **canonicalize everything** (**especially redirects**)

- **Wikipedia's coverage is extensive** (**and growing**)
  - **extrinsic quantity → quality**
  - ***not* intrinsic quality** (**main differentiator**)

    — **pre-Wikipedia (Koningstein et al., 2003–4),**

    ESA (2007), Wiki-linking (Milne and Witten, 2008), etc.

http://wikipapers.referata.com/wiki/List_of_datasets

Football: Forward

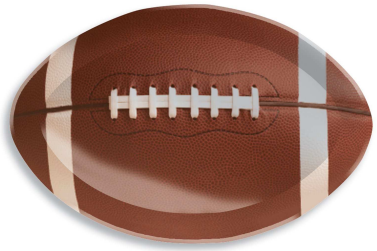
Football: Forward

- **44,984 — Association football**

Football: Forward

- **44,984 — Association football**

- **23,373 — American football**



Football: Back

Football: Back

- **Association football**

Football: Back

- **Association football**
  - **soccer**

Football: Back

- **Association football**
  - **soccer**
  - **association football**

Football: Back

- **Association football**
  - **soccer**
  - **association football**
  - **fútbol**
  - **futbol**
  - **Fußball**
  - **futebol**

Football: Back

- **Association football**
  - **soccer**
  - **association football**
  - **fútbol**
  - **futbol**
  - **Fußball**
  - **futebol**
- **American football**

Football: Back

- **Association football**
  - **soccer**
  - **association football**
  - **fútbol**
  - **futbol**
  - **Fußball**
  - **futebol**
- **American football**
  - **American football**

Football: Back

- **Association football**
  - **soccer**
  - **association football**
  - **fútbol**
  - **futbol**
  - **Fußball**
  - **futebol**
- **American football**
  - **American football**
  - **fútbol americano**

Football: Back

- **Association football**
  - **soccer**
  - **association football**
  - **fútbol**
  - **futbol**
  - **Fußball**
  - **futebol**
- **American football**
  - **American football**
  - **fútbol americano**
  - **football américain**

Named Entities: Highly Ambiguous

Named Entities: Highly Ambiguous

- **people named after other people**

Named Entities: Highly Ambiguous

- **people named after other people**
- **places named after other places**

Named Entities: Highly Ambiguous

- **people named after other people**
- **places named after other places**
- **people named after places where they are from**

Named Entities: Highly Ambiguous

- **people named after other people**
- **places named after other places**
- **people named after places where they are from**
- **places named after people who founded them**

Named Entities: Highly Ambiguous

- **people** named after other **people**
- **places** named after other **places**
- **people** named after **places** where they are from
- **places** named after **people** who founded them
- **organizations** named after **people** or **places**

Named Entities: Highly Ambiguous

- people named after other people
- places named after other places
- people named after places where they are from
- places named after people who founded them
- organizations named after people or places
- organizations become places...

Named Entities: Example — Stanford

Named Entities: Example — Stanford

1.	**Stanford University**	**50.3 ORG**

Named Entities: Example — Stanford

1.	**Stanford University**	**50.3**	**ORG**
2.	**Stanford (disambiguation)**	**7.7**	**—**

Named Entities: Example — Stanford

1.	**Stanford University**	**50.3**	**ORG**
2.	**Stanford (disambiguation)**	**7.7**	**—**
3.	**Stanford, California**	**7.5**	**LOC**

Named Entities: Example — Stanford

1.	**Stanford University**	**50.3**	**ORG**
2.	**Stanford (disambiguation)**	**7.7**	**—**
3.	**Stanford, California**	**7.5**	**LOC**
4.	**Stanford Cardinal football**	**5.7**	**ORG**

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—
8.	Stanford, Kentucky	1.7	LOC

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—
8.	Stanford, Kentucky	1.7	LOC
9.	Stanford, Norfolk	1.0	LOC

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—
8.	Stanford, Kentucky	1.7	LOC
9.	Stanford, Norfolk	1.0	LOC
10.	Bank of the West Classic	1.0	—

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—
8.	Stanford, Kentucky	1.7	LOC
9.	Stanford, Norfolk	1.0	LOC
10.	Bank of the West Classic	1.0	—
11.	Stanford, Illinois	0.9	LOC

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—
8.	Stanford, Kentucky	1.7	LOC
9.	Stanford, Norfolk	1.0	LOC
10.	Bank of the West Classic	1.0	—
11.	Stanford, Illinois	0.9	LOC
12.	Leland Stanford	0.9	PER

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—
8.	Stanford, Kentucky	1.7	LOC
9.	Stanford, Norfolk	1.0	LOC
10.	Bank of the West Classic	1.0	—
11.	Stanford, Illinois	0.9	LOC
12.	Leland Stanford	0.9	PER
13.	Charles Villiers Stanford	0.8	PER

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—
8.	Stanford, Kentucky	1.7	LOC
9.	Stanford, Norfolk	1.0	LOC
10.	Bank of the West Classic	1.0	—
11.	Stanford, Illinois	0.9	LOC
12.	Leland Stanford	0.9	PER
13.	Charles Villiers Stanford	0.8	PER
14.	Stanford, New York	0.8	LOC

Named Entities: Example — Stanford

1.	Stanford University	50.3	ORG
2.	Stanford (disambiguation)	7.7	—
3.	Stanford, California	7.5	LOC
4.	Stanford Cardinal football	5.7	ORG
5.	Stanford Cardinal	4.1	—
6.	Stanford Cardinal men's basketball	2.0	ORG
7.	Stanford prison experiment	2.0	—
8.	Stanford, Kentucky	1.7	LOC
9.	Stanford, Norfolk	1.0	LOC
10.	Bank of the West Classic	1.0	—
11.	Stanford, Illinois	0.9	LOC
12.	Leland Stanford	0.9	PER
13.	Charles Villiers Stanford	0.8	PER
14.	Stanford, New York	0.8	LOC
15.	Stanford, Bedfordshire	0.8	LOC

Named Entities: Objective Evaluation

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**
  - **no learning, ignores context, not language-specific...**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**
  - **no learning, ignores context, not language-specific...**
  - **beats the median entry in all competitions! (so far)**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**
  - **no learning, ignores context, not language-specific...**
  - **beats the median entry in all competitions! (so far)**
  - **tops most entries with a simple additional heuristic**

(Chang et al., 2010)

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**
  - **no learning, ignores context, not language-specific...**
  - **beats the median entry in all competitions! (so far)**
  - **tops most entries with a simple additional heuristic**

(Chang et al., 2010)

- **abstract away sheer engineering effort**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**
  - **no learning, ignores context, not language-specific...**
  - **beats the median entry in all competitions! (so far)**
  - **tops most entries with a simple additional heuristic**

(Chang et al., 2010)

- **abstract away sheer engineering effort**
  - **let research focus on context-sensitive techniques**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**
  - **no learning, ignores context, not language-specific...**
  - **beats the median entry in all competitions! (so far)**
  - **tops most entries with a simple additional heuristic**

(Chang et al., 2010)

- **abstract away sheer engineering effort**
  - **let research focus on context-sensitive techniques**
  - **machine learning**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**
  - **no learning, ignores context, not language-specific...**
  - **beats the median entry in all competitions! (so far)**
  - **tops most entries with a simple additional heuristic**

(Chang et al., 2010)

- **abstract away sheer engineering effort**
  - **let research focus on context-sensitive techniques**
  - **machine learning, linguistic features**

Named Entities: Objective Evaluation

- **entity linking** **(TAC-KBP)**
  - **task: disambiguate entity mentions in text, by linking to appropriate Wikipedia article — e.g., George Bush junior versus senior...**

- **dictionary baseline: simple look-ups (as MFS in WSD)**
  - **return highest scoring concept for every string mention**
  - **no learning, ignores context, not language-specific...**
  - **beats the median entry in all competitions! (so far)**
  - **tops most entries with a simple additional heuristic**

(Chang et al., 2010)

- **abstract away sheer engineering effort**
  - **let research focus on context-sensitive techniques**
  - **machine learning, linguistic features, etc.**

From Words to Concepts and Back:

Examples:

- word sense disambiguation
- named entity recognition

From Words to Concepts and Back:

Examples:

- word sense disambiguation
- named entity recognition
- entity linking

From Words to Concepts and Back:

Examples:

- word sense disambiguation
- named entity recognition
- entity linking
- coreference resolution

From Words to Concepts and Back:

Examples:

- word sense disambiguation
- named entity recognition
- entity linking
- coreference resolution
- web search

From Words to Concepts and Back:

Examples (Recognition):

- word sense disambiguation
- named entity recognition
- entity linking
- coreference resolution
- web search

From Words to Concepts and Back:

— inverse problem —

Examples (Generation):

From Words to Concepts and Back:

— inverse problem —

Examples (Generation):

- word synonyms

From Words to Concepts and Back:

— inverse problem —

Examples (Generation):

- word synonyms
- paraphrasing

From Words to Concepts and Back:

— inverse problem —

Examples (Generation):

- word synonyms
- paraphrasing
- summarization

From Words to Concepts and Back:

— inverse problem —

Examples (Generation):

- word synonyms
- paraphrasing
- summarization
- translation

From Words to Concepts and Back:

— inverse problem —

Examples (Generation):

- word synonyms
- paraphrasing
- summarization
- translation
- keyword targeting

From Words to Concepts and Back:

Comes up in IR and NLP all the time!

From Words to Concepts and Back:

Comes up in IR and NLP all the time!

Good engineering:

From Words to Concepts and Back:

Comes up in IR and NLP all the time!

Good engineering: modularity and abstraction.

From Words to Concepts and Back:

Comes up in IR and NLP all the time!

Good engineering: modularity and abstraction.

- **Dictionary modules: stubs.**

From Words to Concepts and Back:

Comes up in IR and NLP all the time!

Good engineering: modularity and abstraction.

- **Dictionary modules: stubs.**
- **Interface is conditional probabilities:**

From Words to Concepts and Back:

Comes up in IR and NLP all the time!

Good engineering: modularity and abstraction.

- **Dictionary modules**: stubs.
- **Interface is conditional probabilities**:
  — $\mathbb{P}(\text{concept} \mid \text{words})$;

From Words to Concepts and Back:

Comes up in IR and NLP all the time!

Good engineering: modularity and abstraction.

- **Dictionary modules: stubs.**
- **Interface is conditional probabilities:**
  — $\mathbb{P}(\text{concept} \mid \text{words})$; and $\mathbb{P}(\text{words} \mid \text{concept})$.

From Words to Concepts and Back:

Comes up in IR and NLP all the time!

Good engineering: modularity and abstraction.

- **Dictionary modules**: stubs.
- **Interface is conditional probabilities**:
  — $\mathbb{P}(\text{concept} \mid \text{words})$; and $\mathbb{P}(\text{words} \mid \text{concept})$.

Conceptually trivial platform (hides engineering/systems details).

Another Example: — Soft_drink

Another Example: — Soft_drink

- **Normalized** (for capitalization, pluralization and punctuation differences).

Another Example: — Soft_drink

- **Normalized** (for capitalization, pluralization and punctuation differences).

1.	soft drink	28.6
2.	soda	5.5
3.	soda pop	0.9
4.	fizzy drinks	0.6
5.	carbonated beverages	0.3
6.	non-alcoholic	0.2
7.	soft	0.1
8.	pop	0.1
9.	carbonated soft drink	0.1
10.	aerated water	0.1

Another Example: — Soft_drink

- **Normalized** (for capitalization, pluralization and punctuation differences).

1.	soft drink	28.6
2.	soda	5.5
3.	soda pop	0.9
4.	fizzy drinks	0.6
5.	carbonated beverages	0.3
6.	non-alcoholic	0.2
7.	soft	0.1
8.	pop	0.1
9.	carbonated soft drink	0.1
10.	aerated water	0.1

- **Restricted to English Wikipedia** (and hence missing 2/3 of the data).

WYSIWYG Examples: — see paper and data

WYSIWYG Examples: — see paper and data

- **A small, manageable one:** $s =$ **bushbabies:**

WYSIWYG Examples: — see paper and data

- **A small, manageable one:** $s =$ **bushbabies**:

$\hat{\mathbb{P}}(\textbf{URL} \mid s)$	**URL**	*(and Associated Scores)*
0.966102	Galago	**D W:110/111 W08 W09 WDB w:2/5 w':2/2**
0.0169492	bushbaby	**w:2/5**
0.00847458	Lesser_bushbaby	**W:1/111 W08 W09 WDB**
0.00847458	bushbabies	**c t w:1/5**

WYSIWYG Examples: — see paper and data

- **A small, manageable one:** $s =$ **bushbabies:**

$\hat{\mathbb{P}}(\textbf{URL} \mid s)$	**URL**	*(and Associated Scores)*
0.966102	`Galago`	**D W:110/111 W08 W09 WDB w:2/5 w':2/2**
0.0169492	`bushbaby`	**w:2/5**
0.00847458	`Lesser_bushbaby`	**W:1/111 W08 W09 WDB**
0.00847458	`bushbabies`	**c t w:1/5**

- `README` **file has (much) more about the features;**

WYSIWYG Examples: — see paper and data

- **A small, manageable one:** $s =$ **bushbabies:**

$\hat{\mathbb{P}}(\textbf{URL} \mid s)$	**URL**	*(and Associated Scores)*
0.966102	`Galago`	**D W:110/111 W08 W09 WDB w:2/5 w':2/2**
0.0169492	`bushbaby`	**w:2/5**
0.00847458	`Lesser_bushbaby`	**W:1/111 W08 W09 WDB**
0.00847458	`bushbabies`	**c t w:1/5**

- `README` **file has (much) more about the features;**
- **More than half the paper is detailed examples...**

Resource:

Resource:

- **noisy unfiltered cross-lingual dictionary designed for recall**

Resource:

- **noisy unfiltered cross-lingual dictionary designed for recall**
  - **e.g., "click here" or "on Wikipedia"**

Resource:

- **noisy unfiltered cross-lingual dictionary designed for recall**
  - **e.g., "click here" or "on Wikipedia"**
  - **reconciliation of canonical URLs for non-existent pages**

Resource:

- **noisy unfiltered cross-lingual dictionary designed for recall**
  - **e.g., "click here" or "on Wikipedia"**
  - **reconciliation of canonical URLs for non-existent pages**
  - **contradictory redirects** (Wikipedia snapshots from different times)

Resource:

- **noisy unfiltered cross-lingual dictionary designed for recall**
  - **e.g., "click here" or "on Wikipedia"**
  - **reconciliation of canonical URLs for non-existent pages**
  - **contradictory redirects (Wikipedia snapshots from different times)**
  - **but... lots of features to help filter out the noise!**

Resource:

- **noisy unfiltered cross-lingual dictionary designed for recall**
  - **e.g., "click here" or "on Wikipedia"**
  - **reconciliation of canonical URLs for non-existent pages**
  - **contradictory redirects (Wikipedia snapshots from different times)**
  - **but... lots of features to help filter out the noise!**
  - **suitable for use with machine learning techniques**

Resource:

- noisy unfiltered cross-lingual dictionary designed for recall
  - e.g., "click here" or "on Wikipedia"
  - reconciliation of canonical URLs for non-existent pages
  - contradictory redirects (Wikipedia snapshots from different times)
  - but... lots of features to help filter out the noise!
  - suitable for use with machine learning techniques

http://nlp.stanford.edu/pubs/crosswikis-data.tar.bz2

Resource:

- **noisy unfiltered cross-lingual dictionary designed for recall**
  - **e.g., "click here" or "on Wikipedia"**
  - **reconciliation of canonical URLs for non-existent pages**
  - **contradictory redirects (Wikipedia snapshots from different times)**
  - **but... lots of features to help filter out the noise!**
  - **suitable for use with machine learning techniques**

http://nlp.stanford.edu/pubs/crosswikis-data.tar.bz2

— earlier work with E. Agirre, E. Yeh, C. Manning and D. Jurafsky

Resource:

- noisy unfiltered cross-lingual dictionary designed for recall
  - e.g., "click here" or "on Wikipedia"
  - reconciliation of canonical URLs for non-existent pages
  - contradictory redirects (Wikipedia snapshots from different times)
  - but... lots of features to help filter out the noise!
  - suitable for use with machine learning techniques

http://nlp.stanford.edu/pubs/crosswikis-data.tar.bz2

— earlier work with E. Agirre, E. Yeh, C. Manning and D. Jurafsky

- cleaner filtered English dictionary, designed for precision

Resource:

- noisy unfiltered cross-lingual dictionary designed for recall
  - e.g., "click here" or "on Wikipedia"
  - reconciliation of canonical URLs for non-existent pages
  - contradictory redirects (Wikipedia snapshots from different times)
  - but... lots of features to help filter out the noise!
  - suitable for use with machine learning techniques

http://nlp.stanford.edu/pubs/crosswikis-data.tar.bz2

— earlier work with E. Agirre, E. Yeh, C. Manning and D. Jurafsky

- cleaner filtered English dictionary, designed for precision

To be released soon!

Resource:

- noisy unfiltered cross-lingual dictionary designed for recall
  - e.g., "click here" or "on Wikipedia"
  - reconciliation of canonical URLs for non-existent pages
  - contradictory redirects (Wikipedia snapshots from different times)
  - but... lots of features to help filter out the noise!
  - suitable for use with machine learning techniques

http://nlp.stanford.edu/pubs/crosswikis-data.tar.bz2

— earlier work with E. Agirre, E. Yeh, C. Manning and D. Jurafsky

- cleaner filtered English dictionary, designed for precision

To be released soon! — by A. Subramanya, S. Singh, F. Pereira and A. McCallum

Resource:

- noisy unfiltered cross-lingual dictionary designed for recall
  - e.g., "click here" or "on Wikipedia"
  - reconciliation of canonical URLs for non-existent pages
  - contradictory redirects (Wikipedia snapshots from different times)
  - but... lots of features to help filter out the noise!
  - suitable for use with machine learning techniques

http://nlp.stanford.edu/pubs/crosswikis-data.tar.bz2

— earlier work with E. Agirre, E. Yeh, C. Manning and D. Jurafsky

- cleaner filtered English dictionary, designed for precision

To be released soon! — by A. Subramanya, S. Singh, F. Pereira and A. McCallum

We hope you will find creative uses for these! :)

Thanks!

Yet in each word some concept there must be...

Quite true! But don't torment yourself too anxiously;
For at the point where concepts fail,
At the right time a word is thrust in there.

— Mephistopheles, in Goethe's *Faust* (Part I, Scene III, as translated by G.M. Priest)

http://www.levity.com/alchemy/faust05.html

Thanks!

Any questions?