Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova (angelina@ifi.uio.no)

Language Technology Group, Department of Informatics, University of Oslo

Goals

- Develop dictionaries from the Wikipedia link structure
- Evaluate quality and examine the content
- Apply the dictionary to statistical machine translation

Motivation

- Bilingual dictionaries are important resources for NLP
- Manual development is expensive
- Internet-resources provide terminology which is not present in traditional dictionaries

Features of Wikipedia

Entity page is an article that contains information focused on one single entity, such as a person, a place, or a work of art.

Redirect page is a page without article body that can provide:
- equivalent names for an entity
- Federation of Czechoslovakia → Czechoslovakia
- spelling resolution
- Microsoft’s → Microsoft
- spelling correction
- Wikipedia → Wikipedia

Interlanguage links is a link from an article to a presumably equivalent article in another language.

Example of the Wiki-dictionary entry

Czechoslovakia — Federation of Czechoslovakia — Czechoslovakia — Czechoslovakia — CSFR

Example of translation performed by MT system trained without and with the Wiki-dictionary

Without Wiki-dictionary: sociologist at yale university immanuel sannepeham believes that by 2050 , lenin inevitably become a national hero russia . With Wiki-dictionary: marketing sociology at yale university , immanuel vallierstein believes that by 2050 , lenin inevitably will be the national hero russia .

Evaluation tests

UMC is a test set from the sentence-aligned ÚFAL Multilingual Corpora, a parallel corpus of texts in Czech, Russian and English languages.

Wiki-set is a set of English sentences collected from three Wikipedia articles in the versions that were available on July 19, 2011.

Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova (angelina@ifi.uio.no)

Language Technology Group, Department of Informatics, University of Oslo

Goals

- Develop dictionaries from the Wikipedia link structure
- Evaluate quality and examine the content
- Apply the dictionary to statistical machine translation

Motivation

- Bilingual dictionaries are important resources for NLP
- Manual development is expensive
- Internet-resources provide terminology which is not present in traditional dictionaries

Features of Wikipedia

Entity page is an article that contains information focused on one single entity, such as a person, a place, or a work of art.

Redirect page is a page without article body that can provide:
- equivalent names for an entity
- Federation of Czechoslovakia → Czechoslovakia
- spelling resolution
- Microsoft’s → Microsoft
- spelling correction
- Wikipedia → Wikipedia

Interlanguage links is a link from an article to a presumably equivalent article in another language.

Example of the Wiki-dictionary entry

Czechoslovakia — Federation of Czechoslovakia — Czechoslovakia — Czechoslovakia — CSFR

Example of translation performed by MT system trained without and with the Wiki-dictionary

Without Wiki-dictionary: sociologist at yale university immanuel sannepeham believes that by 2050 , lenin inevitably become a national hero russia . With Wiki-dictionary: marketing sociology at yale university , immanuel vallierstein believes that by 2050 , lenin inevitably will be the national hero russia .

Evaluation tests

UMC is a test set from the sentence-aligned ÚFAL Multilingual Corpora, a parallel corpus of texts in Czech, Russian and English languages.

Wiki-set is a set of English sentences collected from three Wikipedia articles in the versions that were available on July 19, 2011.

Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova (angelina@ifi.uio.no)

Language Technology Group, Department of Informatics, University of Oslo

Goals

- Develop dictionaries from the Wikipedia link structure
- Evaluate quality and examine the content
- Apply the dictionary to statistical machine translation

Motivation

- Bilingual dictionaries are important resources for NLP
- Manual development is expensive
- Internet-resources provide terminology which is not present in traditional dictionaries

Features of Wikipedia

Entity page is an article that contains information focused on one single entity, such as a person, a place, or a work of art.

Redirect page is a page without article body that can provide:
- equivalent names for an entity
- Federation of Czechoslovakia → Czechoslovakia
- spelling resolution
- Microsoft’s → Microsoft
- spelling correction
- Wikipedia → Wikipedia

Interlanguage links is a link from an article to a presumably equivalent article in another language.

Example of the Wiki-dictionary entry

Czechoslovakia — Federation of Czechoslovakia — Czechoslovakia — Czechoslovakia — CSFR

Example of translation performed by MT system trained without and with the Wiki-dictionary

Without Wiki-dictionary: sociologist at yale university immanuel sannepeham believes that by 2050 , lenin inevitably become a national hero russia . With Wiki-dictionary: marketing sociology at yale university , immanuel vallierstein believes that by 2050 , lenin inevitably will be the national hero russia .

Evaluation tests

UMC is a test set from the sentence-aligned ÚFAL Multilingual Corpora, a parallel corpus of texts in Czech, Russian and English languages.

Wiki-set is a set of English sentences collected from three Wikipedia articles in the versions that were available on July 19, 2011.

Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova (angelina@ifi.uio.no)

Language Technology Group, Department of Informatics, University of Oslo

Goals

- Develop dictionaries from the Wikipedia link structure
- Evaluate quality and examine the content
- Apply the dictionary to statistical machine translation

Motivation

- Bilingual dictionaries are important resources for NLP
- Manual development is expensive
- Internet-resources provide terminology which is not present in traditional dictionaries

Features of Wikipedia

Entity page is an article that contains information focused on one single entity, such as a person, a place, or a work of art.

Redirect page is a page without article body that can provide:
- equivalent names for an entity
- Federation of Czechoslovakia → Czechoslovakia
- spelling resolution
- Microsoft’s → Microsoft
- spelling correction
- Wikipedia → Wikipedia

Interlanguage links is a link from an article to a presumably equivalent article in another language.

Example of the Wiki-dictionary entry

Czechoslovakia — Federation of Czechoslovakia — Czechoslovakia — Czechoslovakia — CSFR

Example of translation performed by MT system trained without and with the Wiki-dictionary

Without Wiki-dictionary: sociologist at yale university immanuel sannepeham believes that by 2050 , lenin inevitably become a national hero russia . With Wiki-dictionary: marketing sociology at yale university , immanuel vallierstein believes that by 2050 , lenin inevitably will be the national hero russia .

Evaluation tests

UMC is a test set from the sentence-aligned ÚFAL Multilingual Corpora, a parallel corpus of texts in Czech, Russian and English languages.

Wiki-set is a set of English sentences collected from three Wikipedia articles in the versions that were available on July 19, 2011.

Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova (angelina@ifi.uio.no)

Language Technology Group, Department of Informatics, University of Oslo

Goals

- Develop dictionaries from the Wikipedia link structure
- Evaluate quality and examine the content
- Apply the dictionary to statistical machine translation

Motivation

- Bilingual dictionaries are important resources for NLP
- Manual development is expensive
- Internet-resources provide terminology which is not present in traditional dictionaries

Features of Wikipedia

Entity page is an article that contains information focused on one single entity, such as a person, a place, or a work of art.

Redirect page is a page without article body that can provide:
- equivalent names for an entity
- Federation of Czechoslovakia → Czechoslovakia
- spelling resolution
- Microsoft’s → Microsoft
- spelling correction
- Wikipedia → Wikipedia

Interlanguage links is a link from an article to a presumably equivalent article in another language.

Example of the Wiki-dictionary entry

Czechoslovakia — Federation of Czechoslovakia — Czechoslovakia — Czechoslovakia — CSFR

Example of translation performed by MT system trained without and with the Wiki-dictionary

Without Wiki-dictionary: sociologist at yale university immanuel sannepeham believes that by 2050 , lenin inevitably become a national hero russia . With Wiki-dictionary: marketing sociology at yale university , immanuel vallierstein believes that by 2050 , lenin inevitably will be the national hero russia .

Evaluation tests

UMC is a test set from the sentence-aligned ÚFAL Multilingual Corpora, a parallel corpus of texts in Czech, Russian and English languages.

Wiki-set is a set of English sentences collected from three Wikipedia articles in the versions that were available on July 19, 2011.

Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova (angelina@ifi.uio.no)

Language Technology Group, Department of Informatics, University of Oslo

Goals

- Develop dictionaries from the Wikipedia link structure
- Evaluate quality and examine the content
- Apply the dictionary to statistical machine translation

Motivation

- Bilingual dictionaries are important resources for NLP
- Manual development is expensive
- Internet-resources provide terminology which is not present in traditional dictionaries

Features of Wikipedia

Entity page is an article that contains information focused on one single entity, such as a person, a place, or a work of art.

Redirect page is a page without article body that can provide:
- equivalent names for an entity
- Federation of Czechoslovakia → Czechoslovakia
- spelling resolution
- Microsoft’s → Microsoft
- spelling correction
- Wikipedia → Wikipedia

Interlanguage links is a link from an article to a presumably equivalent article in another language.

Example of the Wiki-dictionary entry

Czechoslovakia — Federation of Czechoslovakia — Czechoslovakia — Czechoslovakia — CSFR

Example of translation performed by MT system trained without and with the Wiki-dictionary

Without Wiki-dictionary: sociologist at yale university immanuel sannepeham believes that by 2050 , lenin inevitably become a national hero russia . With Wiki-dictionary: marketing sociology at yale university , immanuel vallierstein believes that by 2050 , lenin inevitably will be the national hero russia .

Evaluation tests

UMC is a test set from the sentence-aligned ÚFAL Multilingual Corpora, a parallel corpus of texts in Czech, Russian and English languages.

Wiki-set is a set of English sentences collected from three Wikipedia articles in the versions that were available on July 19, 2011.

Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova (angelina@ifi.uio.no)

Language Technology Group, Department of Informatics, University of Oslo

Goals

- Develop dictionaries from the Wikipedia link structure
- Evaluate quality and examine the content
- Apply the dictionary to statistical machine translation

Motivation

- Bilingual dictionaries are important resources for NLP
- Manual development is expensive
- Internet-resources provide terminology which is not present in traditional dictionaries

Features of Wikipedia

Entity page is an article that contains information focused on one single entity, such as a person, a place, or a work of art.

Redirect page is a page without article body that can provide:
- equivalent names for an entity
- Federation of Czechoslovakia → Czechoslovakia
- spelling resolution
- Microsoft’s → Microsoft
- spelling correction
- Wikipedia → Wikipedia

Interlanguage links is a link from an article to a presumably equivalent article in another language.

Example of the Wiki-dictionary entry

Czechoslovakia — Federation of Czechoslovakia — Czechoslovakia — Czechoslovakia — CSFR

Example of translation performed by MT system trained without and with the Wiki-dictionary

Without Wiki-dictionary: sociologist at yale university immanuel sannepeham believes that by 2050 , lenin inevitably become a national hero russia . With Wiki-dictionary: marketing sociology at yale university , immanuel vallierstein believes that by 2050 , lenin inevitably will be the national hero russia .

Evaluation tests

UMC is a test set from the sentence-aligned ÚFAL Multilingual Corpora, a parallel corpus of texts in Czech, Russian and English languages.

Wiki-set is a set of English sentences collected from three Wikipedia articles in the versions that were available on July 19, 2011.