Aligning Sense Inventories in Wikipedia and WordNet

TECHNISCHE UNIVERSITÄT DARMSTADT

Elisabeth Wolf and Iryna Gurevych
Ubiquitous Knowledge Processing (UKP) Lab
Technische Universität Darmstadt, Germany
May 19th, 2010

Motivation

Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

Motivation

Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

✓ precise taxonomy

textual information

size

Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

✓ precise taxonomy

✗ textual information

✗ size

Motivation

Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

✓ precise taxonomy	✗
✗ textual information	✓
✗ size	✓

Motivation

Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

✓ precise taxonomy	✗
✗ textual information	✓
✗ size	✓
multilingual	✓

Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

✓ precise taxonomy	✓
✓ textual information	✓
✓ size	✓
multilingual	✓

Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Two main benefits:

1. enhanced sense representation
2. increase of sense coverage

Alignment on sense level

WordNet

- precise taxonomy
- textual information
- size
- multilingual

Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

WordNet

- S: (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- S: (n) **damper**, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) **damper** (a depressing restraint) *"rain put a damper on our picnic plans"*

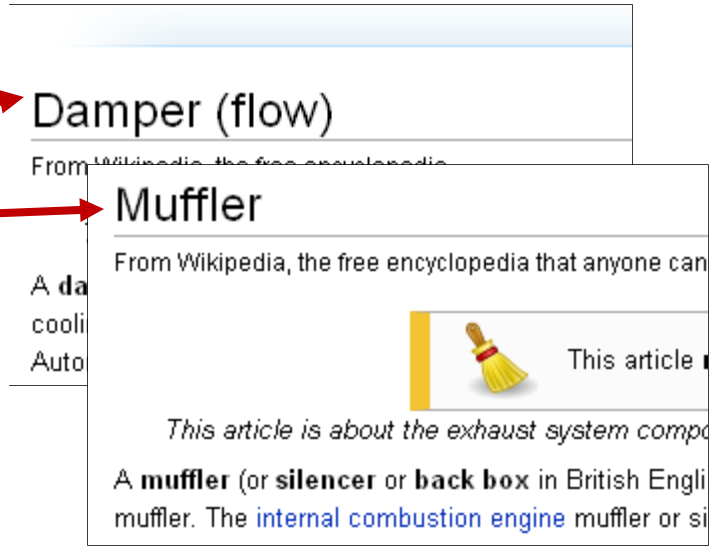
Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

WordNet

WIKIPEDIA
The Free Encyclopedia

- S: (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- S: (n) **damper**, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) **damper** (a depressing restraint) *"rain put a damper on our picnic plans"*



Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

WordNet

WIKIPEDIA
The Free Encyclopedia

- S: (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- S: (n) **damper**, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) **damper** (a depressing restraint) *"rain put a damper on our picnic plans"*

Damper (flow)

From

A da

cooli

Auto

Muffler

From Wikipedia, the free encyclopedia that anyone can

This article

This article is about the exhaust system compo

A **muffler** (or **silencer** or **back box** in British Engli

muffler. The internal combustion engine muffler or si

Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

WordNet

WIKIPEDIA
The Free Encyclopedia

- S: (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- S: (n) **damper**, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) **damper** (a depressing restraint) *"rain put a damper on our picnic plans"*

Damper (flow)

From

Muffler

A da

cooli

Auto

Damper (food)

From Wikipedia, the free encyclopedia

For other uses of the term "damper", see Damp

Damper is a traditional Australian soda bread prep
Australian dish. It is also made in camping situatior

Damper was originally developed by stockmen who

Related Work

Alignment of Wikipedia and WordNet

TECHNISCHE UNIVERSITÄT DARMSTADT

- Main focus on Wikipedia's **category** system and **infoboxes**
  - goal: semantically enriched ontology
  - e.g., Suchanek et al., 2007, Toral et al., 2008, 2009, Ponzetto and Navigli, 2009

- Two works consider Wikipedia's **articles** and WordNet
  - Ruiz-Casado et al., 2005: Simple English Wikipedia
    - automatic mapping based on string comparison methods
    - no analysis regarding the complementarity of senses
  - Mihalcea, 2007: Wikipedia as a source of sense annotations
    - manually mapped Wikipedia articles to WordNet synsets

Related Work

Alignment of Wikipedia and WordNet

TECHNISCHE UNIVERSITÄT DARMSTADT

- Main focus on Wikipedia's **category** system and **infoboxes**
  - goal: semantically enriched ontology
  - e.g., Suchanek et al., 2007, Toral et al., 2008, 2009, Ponzetto and Navigli, 2009

Our contributions:
- Analysis of sense coverage / complementarity
- Dataset with human annotations
- Insights for the (fully) automatic alignment

- Two w
  - Ruiz
    - automatic mapping based on string comparison methods
    - no analysis regarding the complementarity of senses
  - Mihalcea, 2007: Wikipedia as a source of sense annotations
    - manually mapped Wikipedia articles to WordNet synsets

How to build an aligned sense inventory?

Two-step Approach

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Extraction of potential sense pairs
2. Disambiguation of these pairs

Synset	Gloss + Examples	Article Title	First Paragraph (shortened)
doctor, medico, doc, MD, physician, Dr.	*a licensed medical practitioner; "I felt so bad I went to see my doctor"*	*Physician*	*A physician, medical practitioner, doctor of medicine, or medical doctor practices medicine, and is concerned with [· · ·]*
doctor, medico, doc, MD, physician, Dr.	*a licensed medical practitioner; "I felt so bad I went to see my doctor"*	*Doctor (title)*	*Doctor (gen.: doctoris) means teacher in Latin. The word is originally an agentive noun of the verb docre ('to teach'). It has been used continuously as [· · ·]*

WordNet

How to build an aligned sense inventory?

Two-step Approach

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Extraction of potential sense pairs
2. Disambiguation of these pairs

Synset	Gloss + Examples	Article Title	First Paragraph (shortened)
doctor, medico, doc, MD, physician, Dr.	*a licensed medical practitioner; "I felt so bad I went to see my doctor"*	*Physician*	*A physician, medical practitioner, doctor of medicine, or medical doctor practices medicine, and is concerned with [· · ·]*
doctor, medico, doc, MD, physician, Dr.	*a licensed medical practitioner; "I felt so bad I went to see my doctor"*	*Doctor (title)*	*Doctor (gen.: doctoris) means teacher in Latin. The word is originally an agentive noun of the verb docre ('to teach'). It has been used continuously as [· · ·]*

WordNet

WIKIPEDIA
The Free Encyclopedia

How to build an aligned sense inventory?

1. Extraction of potential sense pairs

TECHNISCHE UNIVERSITÄT DARMSTADT

- S: (n) **handwriting**, hand, script (something written by hand) *"she recognized his handwriting"; "his hand was illegible"*

- for each synonymous word extract all
  - articles with title matching word
  - articles with redirect matching word
  - with or without description tag

- examples
  - article *Script (typefaces)*
  - article *Script(comics)*
  - article *Penmanship* has redirect *Handwriting*

How to build an aligned sense inventory?

1. Extraction of potential sense pairs

TECHNISCHE UNIVERSITÄT DARMSTADT

- S: (n) **handwriting**, hand, script (something written by hand) *"she recognized his handwriting"; "his hand was illegible"*

- for each synonymous word extract all
  - articles with title matching word
  - articles with redirect matching word
  - with or without description tag

- examples
  - article *Script (typefaces)*
  - article *Script(comics)*
  - article *Penmanship* has redirect *Handwriting*

# extracted Wikipedia senses	# WordNet senses	
None	12,493	15.2%
One	27,973	34.1%
More than one	41,649	50.7%

How to build an aligned sense inventory?

2. Disambiguation of potential sense pairs: Human annotations

TECHNISCHE UNIVERSITÄT DARMSTADT

- dataset:
  - 38 synsets yield 297 potential sense pairs
  - 7.82 articles per synset on average

- annotation task:
  - four human annotators
  - label either as same sense or not
  - annotations reliable, inter-annotator agreement:
    - Ao = 0.97, multi-kappa = 0.84

How to build an aligned sense inventory?
Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- 21 synsets aligned with at least one article

- 17 synsets not aligned
  - entry in disambiguation page: 4 (ongoing work)
  - article title / redirect did not match synset words: 2
    (already included in the new version)

How to build an aligned sense inventory?
Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- S: (n) **bandwagon** (a popular trend that attracts growing support)

?

Bandwagon effect

From Wikipedia, the free encyclopedia

consider link labels

anyone interested in partaking – most famously t
rld, Leary ultimately joined the bandwagon of "aci
neeting between Kesey and Leary would resolve

```
Leary ultimately joined
the [[Bandwagon effect|
bandwagon]] of "acid
populism" as well.
```

How to build an aligned sense inventory?
Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- 21 synsets aligned with at least one article

- 17 synsets not aligned
  - entry in disambiguation page: 4
  - article title / redirects did not match synset words: 2
  - remaining not in Wikipedia !

Two main benefits:

1. enhanced sense representation
2. increase of sense coverage

Conclusions
Lessons learned

TECHNISCHE UNIVERSITÄT DARMSTADT

- article – synset is an appropriate level of granularity (cf. Mihalcea, 2007)
- two main benefits:
  - enhanced sense representation
  - complementarity of senses

precise taxonomy ✓
textual information ✓
size ✓

- two-step approach: candidate **extraction** and **disambiguation**
  - extraction:
    - consider link labels and process disambiguation pages
  - disambiguation:
    - high-inter annotator agreement motivates automation

Future Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- generate larger dataset for evaluation
- perform automatic alignment
  - text similarity measures
  - ...

- integrate other resources
  - sense inventories, dictionaries
  - e.g. YAGO

Thank you for your attention!

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous Knowledge Processing

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

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

Backup Slides

TECHNISCHE UNIVERSITÄT DARMSTADT

Terminology

TECHNISCHE UNIVERSITÄT DARMSTADT

WordNet

- S: (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)

Dictionary entries. Encyclopedia articles are about a person, or a group, a concept, a place, a thing, an event, etc. In some cases, a word or phrase itself may be an encyclopedic subject, such as Macedonia (terminology) or truthiness; however, articles rarely, if ever, contain several *distinct* definitions or usages of the article title. Articles about the cultural or mathematical

WordNet sense

Wikipedia sense

How to build an aligned sense inventory?

2. Disambiguation of potential sense pairs: Human study

TECHNISCHE UNIVERSITÄT DARMSTADT

- examples:

WordNet	Wikipedia
<configuration, constellation>: an arrangement of parts or elements	**<Configuration (mathematics)>:** In mathematics, especially geometry, a configuration is an
<alignment>: the spatial property possessed by an arrangement or position of things in a straight line or in parallel lines	**<Typographic alignment>**: In typesetting and page layout, alignment or range, is the setting of text flow or image placement …

Why synset and not word sense level?

TECHNISCHE UNIVERSITÄT DARMSTADT

- S: (n) **handwriting**, hand, script (something written by hand) *"she recognized his handwriting"; "his hand was illegible"*

- the only difference is the number of extracted potential Wikipedia articles
- it is more likely that we extract the appropriate Wikipedia article if we consider all synonymous words in a synset

- if we can align on synset level, we can also align on word sense level

Application of an aligned sense inventory

TECHNISCHE UNIVERSITÄT DARMSTADT

- Increase of sense coverage
- Representation contains relational information from WordNet and encyclopaedic information from Wikipedia (multilingual)

- Ability to automatically acquire sense-tagged corpora in a mono- and multilingual fashion
  - For each WordNet synset, the article text of the aligned Wikipedia sense can be automatically extracted similar to Mihalcea, 2007
- Conversion of sense-tagged corpora (e.g. Senseval) to another sense representation
- Use Wikipedia content to extract new sense frequency values

- Enhance Wikipedia articles with synonyms (e.g. as redirects)

How to build an aligned sense inventory?
Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- 21 synsets aligned with at least one article

- 17 synsets not aligned
  - 4 of them as entry in disambiguation page
  - for 2 of them appropriate article could not be extracted
  - remaining synsets not in Wikipedia (one third)

How to build an aligned sense inventory?
Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- S: (n) **bandwagon** (a popular trend that attracts growing support)

?

Bandwagon effect
From Wikipedia, the free encyclopedia

consider link anchors

anyone interested in partaking – most famously t
rld, Leary ultimately joined the bandwagon of "aci
neeting between Kesey and Leary would resolve

```
Leary ultimately joined
the [[Bandwagon effect|
bandwagon]] of "acid
populism" as well.
```

- S: (n) **bandwagon** (a large ornate wagon for carrying a musical band)

?

Bandwagon
From Wikipedia, the free encyclopedia

Bandwagon may refer to:

- a wagon which carries a band of musicians in a parade or for promotional purposes. Other uses

process of disambiguation pages

Subject of our ongoing work