Using Wikipedia for Automatic Word Sense Disambiguation

Rada Mihalcea
(2007)

Rositsa Nedyalkova
24.1.2011

Proseminar „Word Sense Disambiguation“
Stefan Thater

Inhalt

- Einführung
- Wikipedia
- Wikipedia als annotiertes Korpus
- Annotierte Korpora – Aufbau
- Word Sence Disambiguierung
- Experimente und Resultate
- Diskussion
- Fazit

 Ambiguität 

I promise I'll give you a ring tomorrow.

Ich habe dir einen schönen Strauß gekauft.

 Word sence Disambiguierung 

I promise I'll give you a ring tomorrow.

ring (jewel) / **ring (call)**

Ich habe dir einen schönen Strauß gekauft**.**

Strauß (Bukett) / Strauß (Vogel)

😊 Vielfalt an Methoden 😊

- Überwachte Lernsysteme
- Generierung von annotierten Daten durch eindeutige Synonyme
- Disambiguierung durch Übersetzung in andere Sprache

☺ Die Lösung durch Wikipedia ☺

…und mithilfe der freiwilligen Zusammenarbeit im Netz

Warum Wikipedia?

- Wikipedia als Quelle der Annotation
- Links in Wikipedia

  “Henry Barnard, [[United States|American]] [[educationalist]], was born in [[Hartford, Connecticut]]”

- Gute Resultate

Wikipedia

- Freie Enzyklopädie
  - Gemeinschaftliche Leistung
  - „freedom of contribution“ – Quantität und Qualität
  - Verfügbar für mehr als 200 Sprachen

- Aufbau
  - *Grundeintrag - Article / Page*
  - *Hypertext Dokumente*
  - *Jeder Artikel hat eindeutigen Identifikator*

Wikipedia

Flower

From Wikipedia, the free encyclopedia

For other uses, see Flower (disambiguation).

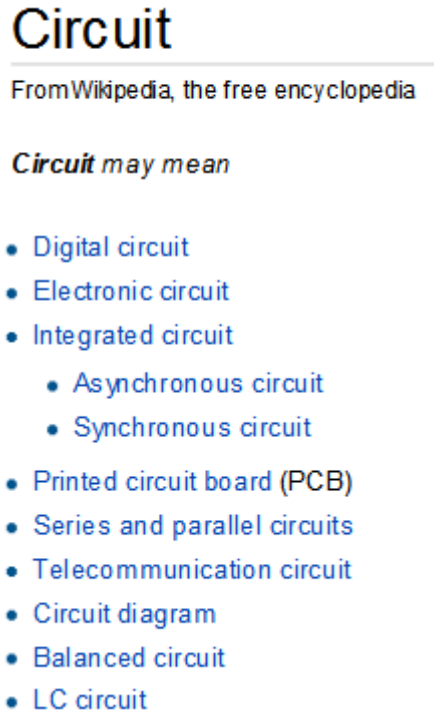
A **flower**, sometimes known as a bloom or blossom, is the reproductive structure found in flowering plants (plants of the division Magnoliophyta, also called angiosperms). The biological function of a flower is to effect reproduction, usually by providing a mechanism for the union of sperm with eggs. Flowers may facilitate outcrossing (fusion of sperm and eggs from different individuals in a population) or allow selfing (fusion of sperm and egg from the same flower). Some flowers produce diaspores without fertilization (parthenocarpy). Flowers contain sporangia and are the site where gametophytes develop. Flowers give rise to fruit and seeds. Many flowers have evolved to be attractive to animals, so as to cause them to be vectors for the transfer of pollen.

Contents [hide]

A poster with twelve species of flowers or clusters of flowers of different families

Wikipedia

- Link
- Piped link
- Redirect page
- Disambiguation page



Circuit

From Wikipedia, the free encyclopedia

***Circuit** may mean*

- Digital circuit
- Electronic circuit
- Integrated circuit
  - Asynchronous circuit
  - Synchronous circuit
- Printed circuit board (PCB)
- Series and parallel circuits
- Telecommunication circuit
- Circuit diagram
- Balanced circuit
- LC circuit

A '''flower''', sometimes known as a [[bloom]] or [[blossom]], is the [[reproduction|reproductive]] structure found in [[flowering plant]]s (plants of the division [[Magnoliophyta]], also called angiosperms).

Flower (disambiguation)

From Wikipedia, the free encyclopedia

A **flower** or **flora** is a reproductive structure found in many plants.

Flower or **Flowers** may also refer to:

Music

Bands

- Flower (band), an American indie rock band (1986–1990)
- The Flowers, a Chinese band
- Icehouse (band) or Flowers, an Australian rock band

Albums

- *Flowers* (Ace of Base album)
- *Flowers* (Echo & the Bunnymen album)
- *Flowers* (The Emotions album)
- *Flowers* (The Rolling Stones album)
- *Flowers* (Joan of Arc album)
- *Flower* (Jody Watley album)

Songs

- "flower" (Kumi Koda song)
- "Flower" (Gackt song)
- "Flower" (Liz Phair song)
- "Flower" (Soundgarden song)
- "Flower" (Tomiko Van song)
- Flower (L'Arc-en-Ciel song)
- "A Flower", a song for voice and closed piano by John Cage

Wikipedia als annotiertes Korpus

- Die Rolle der Links als „sense annotations“

  - Fast jedes Konzept wird dadurch explizit zu anderen korrespondierenden Artikeln verbunden

  - Jeder Link wird manuell von Wikipedia User erstellt

  - meistens genau und hinweisend auf den richtigen Artikel

  - Sehr nützlich für ambige Wörter

Wikipedia als annotiertes Korpus

Beispiel

In 1834, Sumner was admitted to the **[[bar (law)|bar]]** at the age of twenty-three, and entered private practice in Boston.
It is danced in 3/4 time (like most waltzes), with the couple turning approx. 180 degrees every **[[bar (music)|bar]]**.
Vehicles of this type may contain expensive audio players, televisions, video players, and **[[bar (counter)|bar]]**s, often with refrigerators.
Jenga is a popular beer in the **[[bar (establishment)|bar]]**s of Thailand.
This is a disturbance on the water surface of a river or estuary, often cause by the presence of a **[[bar (landform)|bar]]** or dune on the riverbed.

Wikipedia als annotiertes Korpus

- Verzichten auf Disambiguation Page

  ➢ nicht alle Auftreten des Wortes werden aufgelistet

“The blues uses a rhythmic scheme of twelve 4/4 [[measure (music)|bars]]”

  ➢Mögliche Inkonsistenz

- Eigene Datensammlung mithilfe von Wordnet

Annotierte Korpora - Aufbau

1. Auswahl aller Absätze, wo das ambige Wort als (Teil-) Link vorkommt

2. Sammeln aller möglichen Labels für das Wort

3. Mapping

Annotierte Korpora - Aufbau

1. Auswahl aller Absätze, wo das ambige Wort als (Teil-) Link vorkommt

   - Jeder Absatz ist etwa 80 Wörter lang

   - Absätze, wo das Wort groß geschrieben ist, werden nicht berücksichtigt

Annotierte Korpora - Aufbau

2. Sammeln aller möglichen Labels für das Wort

- Auswahl der höchstwertigen Komponenten von den Links

  [[musical notation|bar]] → Label → musical notation

- Das Wort kann selbst die Rolle eines Labels übernehmen
  - Wenn der Link einfach ist und nicht mit Disambiguierungsseite verbindet *[[bar]]*

Annotierte Korpora - Aufbau

3. Mapping

- Jedes Label wird zu der entsprechenden WordNet Bedeutung abgebildet

- Relativ wenig Labels , dadurch schneller Prozess

- Parallele Handannotation (zwei bis drei Annotatoren)

Annotierte Korpora - Aufbau

Beispiel

Word sense	Labels in Wikipedia	WordNet definition
bar (establishment)	bar_(establishment), nightclub gay_club, pub	a room or establishment where alcoholic drinks are served over a counter
bar (counter)	bar_(counter)	a counter where you can obtain food or drink
bar (unit)	bar_(unit)	a unit of pressure equal to a million dynes per square centimeter
bar (music)	bar_(music), measure_music musical_notation	musical notation for a repeating pattern of musical beats
bar (law)	bar_association, bar_law law_society_of_upper_canada state_bar_of_california	the body of individuals qualified to practice law in a particular jurisdiction
bar (landform)	bar_(landform)	a submerged (or partly submerged) ridge in a river or along a shore
bar (metal)	bar_metal, pole_(object)	a rigid piece of metal or wood
bar (sports)	gymnastics_uneven_bars, handle_bar	a horizontal rod that serves as a support for gymnasts as they perform exercises
bar (solid)	candy_bar, chocolate_bar	a block of solid substance

Word Sense Disambiguierung

- Das System
  - Modell zum Vorhersagen bei neuen ambigen Wörter
  - Integrierte lokale und thematische Merkmale
- Vorgehensweise
  - Vorverarbeitung
    - Der Text – Tokens und Part-Of-Speech Tags
    - Kollokationen – Schiebefentsterverfahren
  - Extrahieren der Merkmale aus dem Kontext des Wortes
    - PoS des Wortes und seiner „Nachbaren“
    - Lokaler Kontext - drei Wörter vor und nach dem ambigen Wort
    - Verb und Nomen vor und nach dem Wort
    - Globaler Kontext – Liste der bestimmten Schlüsselwörter

Word Sense Disambiguierung

- Problem
  - ➢Systeme mit unterschiedlichen Testdaten, Annotationen, Bedeutungsinventaren und Korpora

- Lösung
  - ➢Evaluirung für WSD im Vergleich von Systemen an gemeinsamen Daten
  - ➢Inventar bestehend aus mehreren Wort-Bedeutung Mappings
  - ➢Korpus mit hand-annotierten Texten aufgeteilt in Training- und Test Korpus

Experimente und Resultate

- Evaluation
  - Teilmenge der ambigen Wörter bei SENSEVAL-2 und SENSEVAL-3 Evaluation
  - Fixiert auf 49 ambige Nomen SENSEVAL-2 (29) , SENSEVAL-3 (20)
  - Wörter mit nur einem Label in Wikipedia werden entfernt (church, detention)
  - Evaluation erfolgt auf 30 Nomen

Experimente und Reslutate

word	#s	#ex	baselines MFS	baselines LeskC	word sense disambig.
argument	2	114	70.17%	73.63%	**89.47%**
arm	3	291	61.85%	69.31%	**84.87%**
atmosphere	3	773	54.33%	56.62%	**71.66%**
bank	3	1074	**97.20%**	**97.20%**	**97.20%**
bar	10	1108	47.38%	68.09%	**83.12%**
chair	3	194	67.57%	65.78%	**80.92%**
channel	5	366	51.09%	52.50%	**71.85%**
circuit	4	327	85.32%	85.62%	**87.15%**
degree	7	849	58.77%	73.05%	**85.98%**
difference	2	24	**75.00%**	**75.00%**	**75.00%**
disc	3	73	52.05%	52.05%	**71.23%**
dyke	2	76	77.63%	82.00%	**89.47%**
fatigue	3	123	66.66%	70.00%	**93.22%**
grip	3	34	44.11%	**77.00%**	70.58%
image	2	84	69.04%	74.50%	**80.28%**
AVERAGE	3.31	316	72.58%	78.02%	**84.65%**

word	#s	#ex	baselines MFS	baselines LeskC	word sense disambig.
material	3	223	**95.51%**	**95.51%**	**95.51%**
mouth	2	409	94.00%	94.00%	**95.35%**
nature	2	392	**98.72%**	**98.72%**	98.21%
paper	5	895	**96.98%**	**96.98%**	**96.98%**
party	3	764	68.06%	68.28%	**75.91%**
performance	2	271	**95.20%**	**95.20%**	**95.20%**
plan	3	83	77.10%	81.00%	**81.92%**
post	5	33	54.54%	**62.50%**	51.51%
restraint	2	9	77.77%	77.77%	77.77%
sense	2	183	**95.10%**	**95.10%**	**95.10%**
shelter	2	17	**94.11%**	**94.11%**	**94.11%**
sort	2	11	81.81%	**90.90%**	**90.90%**
source	3	78	55.12%	81.00%	**92.30%**
spade	3	46	60.86%	**81.50%**	80.43%
stress	3	565	53.27%	54.28%	**86.37%**
AVERAGE	3.31	316	72.58%	78.02%	**84.65%**

MFS = most frequent sense
LeskC = Lesk-corpus

#s Anzahl Bedeutungen
#ex Anzahl Beispiele

Diskussion

- Ist das System zuverlässig?

  - Durchschnittliche Verbesserung Fehlerrate
    - 44 % vergleicht mit MFS Baseline
    - 30% verlgeicht mit Lesk – Korpus Baseline

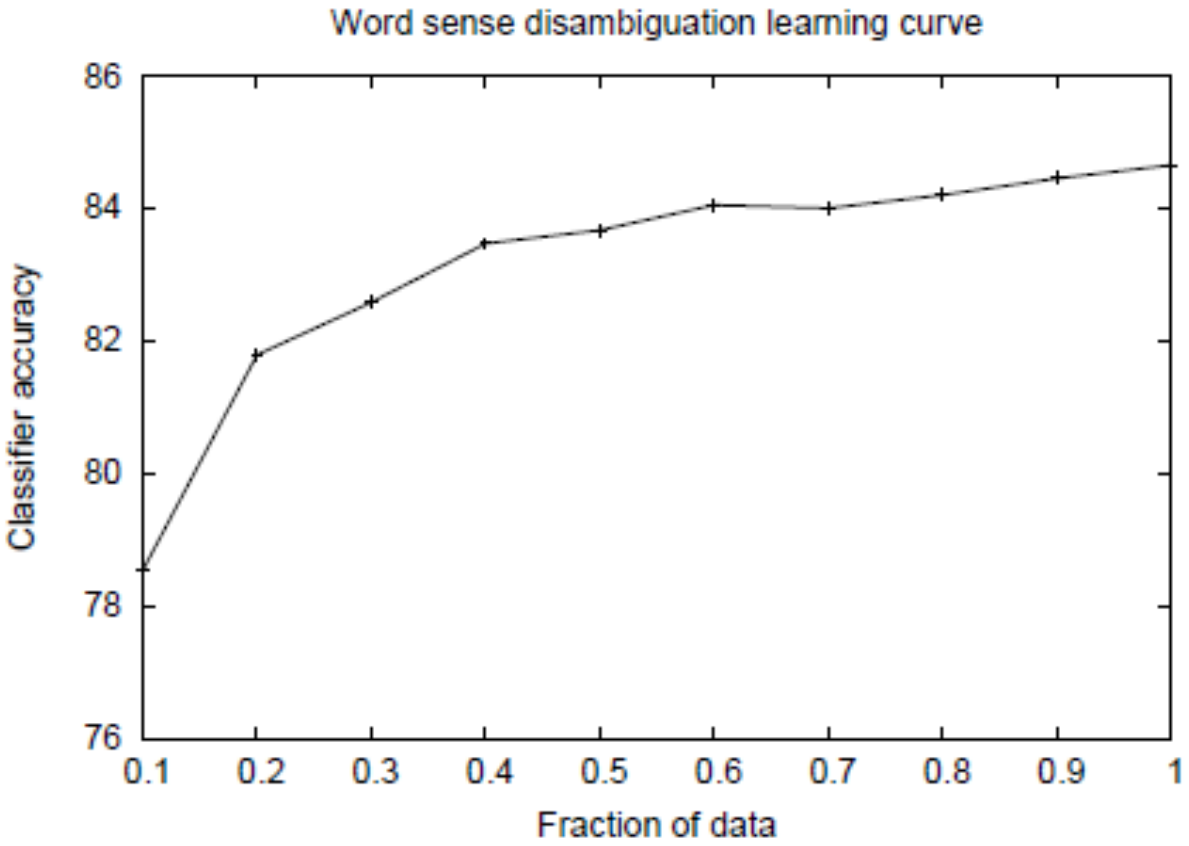
  - Ausnahmen
    - Wörter mit wenig Beispielen in Wikipedia (restraint, shelter) → keine Präzisionsverbesserung
    - Wörter mit ungleichmäßigen Verteilung in Data Set

bank : 1074 Beispiele für Bedeutung „financial institution“ ,

etwa 220 Bespiele für „substance“

Diskussion

- Ist das System zuverlässig?



Diskussion

- Bedeutungsabdeckung – Vergleich mit SENSEVAL

online Kollektion von enzyklopedischen Seiten

vs.

manuell balanciertes British National Korpus

➢ Alle definierte Bedeutungen für den Set von 30 Nomen

➢ Bestimmung des Prozentanteils für jede Bedeutung

chair	sense #1	sense #2	sense #3	sense #4
Wikipedia	68.0 %	31.9 %	0 %	0.1 %
SENSEVAL	87.7 %	6.3 %	6.0 %	0 %

Diskussion

- Korrelationsfaktor zwischen den beiden Korpora

  (r) = 0.51 (mittlere Korrelation)

- SENSEVAL Bedetungsverteilung

  training data ←→ test data

  r = 0.95

Fazit

- Verbesserungen
  - ➢ Einsprachiges Datenmaterial
    - Vermeiden der Einschränkungen mancher Methoden, parallele Texte zu suchen
  - ➢ Die Annotationen befolgen das Zipfsche Gesetz
    - Vermeiden der gleichmäßigen Verteilung von Methoden, die eindeutige Synonyme für die Disambiguierung benutzen
  - ➢ Große Abdeckung
    - Dadurch viel schneller als andere aufgabenorientierte Methoden (Open Mind Word Expert)

Fazit

- Was haben wir gelernt?
  - Wikipedia - gute Quelle für annotierte Training Korpora
  - WSD System mit 30-44% Senkung der Fehlerraten

- Warum lohnt es sich?

1. Die Größe der Datenbank wächst mit jedem Tag
   → damit auch die Menge der annotierten Korpora

2. Wikipedia ist verfügbar für mehr als 200 Sprachen
   → Entwicklung von genauen Word Sense Klassifikatoren für große Anzahl an Sprachen

Fazit

- Was bleibt offen?
  - Niedrigere Fehlerrate..aber im Vergleich zu einfacheren Methoden
  - Definitionen und Annotatitonen meistens nur für Nomen
  - Annotationslabeling - manchmal inkonsistent

 Danke für die Aufmerksamkeit!