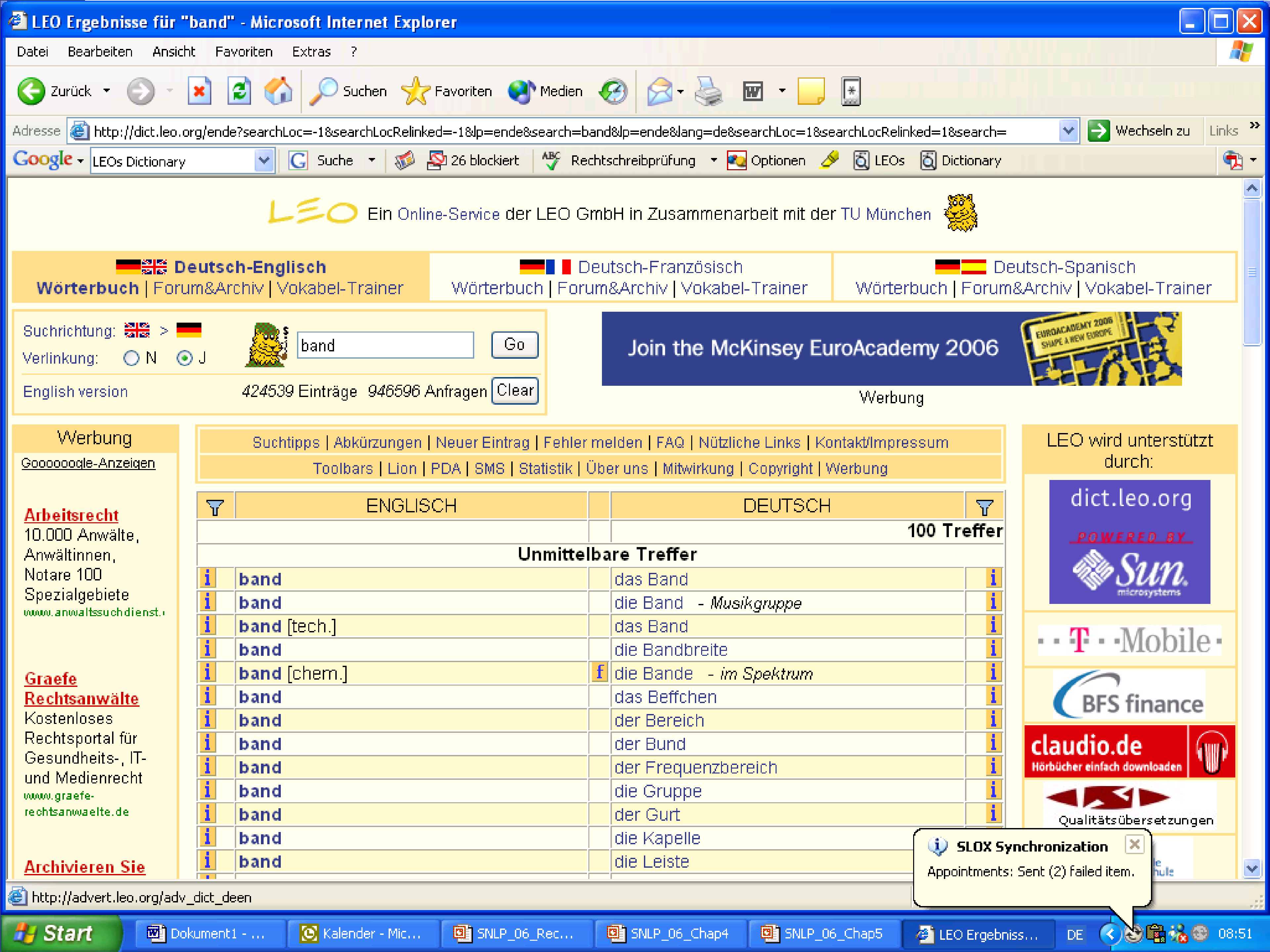
Chapter 7:
Word Sense Disambiguation (WSD)

See: Christopher D. Manning and Hinrich Schütze.
Chapter 7: Word Sense Disambiguation

Introduction and Examples: Word Sense Disambiguation


LEO Ergebnisse für "band" - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten Medien
Adresse http://dict.leo.org/ende?searchLoc=-1&searchLocRelinked=-1&lp=ende&search=band&lp=ende&lang=de&searchLoc=1&searchLocRelinked=1&search=
Wechseln zu Links
LEOs Dictionary Suche 26 blockiert Rechtschreibprüfung Optionen LEOs Dictionary
LEO Ein Online-Service der LEO GmbH in Zusammenarbeit mit der TU München
Deutsch-Englisch
Wörterbuch | Forum&Archiv | Vokabel-Trainer
Deutsch-Französisch
Wörterbuch | Forum&Archiv | Vokabel-Trainer
Deutsch-Spanisch
Wörterbuch | Forum&Archiv | Vokabel-Trainer
Suchrichtung: >
Verlinkung: N J
band
Go
English version
424539 Einträge 946596 Anfragen
Clear
Join the McKinsey EuroAcademy 2006
Werbung
Werbung
Goooooogle-Anzeigen
Arbeitsrecht
10.000 Anwälte, Anwältinnen, Notare 100 Spezialgebiete
www.anwaltssuchdienst.
Graefe Rechtsanwälte
Kostenloses Rechtsportal für Gesundheits-, IT- und Medienrecht
www.graefe-rechtsanwaelte.de
Archivieren Sie
Suchtipps | Abkürzungen | Neuer Eintrag | Fehler melden | FAQ | Nützliche Links | Kontakt/Impressum
Toolbars | Lion | PDA | SMS | Statistik | Über uns | Mitwirkung | Copyright | Werbung
ENGLISCH
DEUTSCH
100 Treffer
Unmittelbare Treffer
band — das Band
band — die Band - Musikgruppe
band [tech.] — das Band
band — die Bandbreite
band [chem.] — die Bande - im Spektrum
band — das Beffchen
band — der Bereich
band — der Bund
band — der Frequenzbereich
band — die Gruppe
band — der Gurt
band — die Kapelle
band — die Leiste
LEO wird unterstützt durch:
dict.leo.org
POWERED BY
Sun microsystems
T-Mobile
BFS finance
claudio.de
Hörbücher einfach downloaden
Qualitätsübersetzungen
SLOX Synchronization
Appointments: Sent (2) failed item.
http://advert.leo.org/adv_dict_deen
Start
Dokument1 - ...
Kalender - Mic...
SNLP_06_Rec...
SNLP_06_Chap4
SNLP_06_Chap5
LEO Ergebniss...
DE
08:51

LEO Ergebnisse für "band" - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Medien

Adresse http://dict.leo.org/ende?searchLoc=-1&searchLocRelinked=-1&lp=ende&search=band&lp=ende&lang=de&searchLoc=1&searchLocRelinked=1&search= Wechseln zu Links

Google LEOs Dictionary Suche 26 blockiert Rechtschreibprüfung Optionen LEOs Dictionary

Archivieren Sie schon?
E-Mail Archivierung ist Pflicht! Geschäftsführer haften Privat.
www.emailarchive.de

Textlink-Werbung:
- Sprachschule Sprachreise
- Preisvergleich Geizhals.at
- Kultur, Film und mehr!
- Nachhilfe - Die Schülerhilfe
- Leiterplatten - Discount
- Preisvergleich
- Lastminute Reisen
- Schönheits-Chirurgie-News
- Wellness bei beauty24.de
- Preissuche
- Preisvergleich
- Lastminute

band	die Leiste
band	die Musikkapelle
band	das Orchester
band	die Schar
band	die Schnur
band [mus.]	der Spielmannszug
band	der Streifen
band	die Truppe
narrowband *also:* narrow-**band** *adj.*	engbandig
narrowband *also:* narrow-**band** *adj.*	schmalbandig
sideband *also:* side **band** [elec.] [telecom.]	das Seitenband
Verben und Verbzusammensetzungen	
to **band** together	sich verbinden
to **band** together	sich vereinigen
to **band** together	sich zusammenrotten
to **band** together	sich zusammentun
to **band** together	zu einer Gruppe vereinigen
to beat the **band**	nie da gewesen sein
to cross-**band** [tech.]	absperren *[Holzverarbeitung]*
Zusammengesetzte Einträge	
abrasive **band** - *cloth* [tech.]	das Bandschleifleinen
abrasive **band** - *paper* [tech.]	das Bandschleifpapier
adhesive **band** [tech.]	das Klischeeklebeband
attenuating **band** [aviat.]	der Dämpfungsbereich
audio **band** [phys.]	der Hörbereich
band aerial	die Bandantenne
band-aid	das Heftpflaster
band-aid *(Amer.)* [med.]	das Pflaster
band-aid *(Amer.)* [med.]	das Wundpflaster
band box	die Hutschachtel

virtuelle hochschule bayern

ÜBERSETZUNGEN Alle Sprachen [DIN]

Hinweis

INFOR MATIK JAHR
Wissenschaftsjahr 2006

Internet

Start Dokument1 - ... Kalender - Mic... SNLP_06_Rec... SNLP_06_Chap4 SNLP_06_Chap5 LEO Ergebniss... DE 08:51

LEO Ergebnisse für "band" - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Medien

Adresse http://dict.leo.org/ende?searchLoc=-1&searchLocRelinked=-1&lp=ende&search=band&lp=ende&lang=de&searchLoc=1&searchLocRelinked=1&search= Wechseln zu Links

Google LEOs Dictionary Suche 26 blockiert Rechtschreibprüfung Optionen LEOs Dictionary

	band ceramics	die Bandkeramik	i
	band collar	der Stehkragen	i
	band-conveyor	das Fließband	i
	band conveyor [tech.]	der Gurtförderer	i
	band-conveyor	das Transportband	i
	band edge	die Bandkante	i
	band emission [autom.]	die Bandemission	i
	band emission [autom.]	die Bandenemission	i
	band gap [phys.]	die Bandlücke	i
	band gate [tech.]	der Bandausschnitt - *Spritzgusswerkzeug [Kunststoffe]*	i
	band grinder [tech.]	die Bandschleifmaschine	i
	band matrix [math.]	die Bandmatrix	i
	band of barrel	das Fassband	i
	band of barrel	der Fassreifen	i
	band of radiation [phys.]	der Strahlungsbereich	i
	band of robbers	die Räuberbande	i
	band overlap [tech.]	die Bandüberlappung	i
	band printer [print.]	der Banddrucker	i
	band radiation [autom.]	die Bandenstrahlung	i
	band resaw [tech.]	die Trennbandsäge	i
	band saw [tech.]	die Bandsäge	i
	band-saw	die Bandsäge	i
	band spectrum [tech.]	das Bandenspektrum	i
	band-spread	die Bandspreizung	i
	band-stand	der Musikpavillon	i
	band structure [phys.]	die Bandstruktur	i
	band-switch	der Bereichsschalter	i
	band-switch	der Bereichsumschalter	i
	band width	die Bandbreite	i

Internet

Start Dokument1 - ... Kalender - Mic... SNLP_06_Rec... SNLP_06_Chap4 SNLP_06_Chap5 LEO Ergebniss... DE 08:51

Word Sense

- Many words have several meanings/senses.
- Consider two senses of the word *bank*
  - The rising ground bordering a lake, river or sea …
  - An establishment for the custody, loan exchange, or issue of money, for the extension of credit, and for facilitating the transmission or funds
- However, the senses are not always so well defined.

Disambiguation

- To determine which of the senses of an ambiguous word is invoked in a particular use of the word.

- **How it is done**:
  - A word is assumed to have a finite number of discrete senses.
  - Look at the context of the word's use to disambiguate.

Senseval web page - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Medien

Adresse http://www.senseval.org/ Wechseln zu Links

Google Senseval Suche 25 blockiert Rechtschreibprüfung Optionen Senseval

SENS EVAL

Evaluation Exercises for the Semantic Analysis of Text ■ Organized by ACL-SIGLEX

CONTENTS

There are now many computer programs for automatically determining the sense of a word in context (Word Sense Disambiguation or WSD). The purpose of Senseval is to evaluate the strengths and weaknesses of such programs with respect to different words, different varieties of language, and different languages.

Senseval-1 took place in the summer of 1998 for English, French, and Italian, culminating in a workshop held at Herstmonceux Castle, Sussex, England on September 2-4.

Senseval-2 took place in the summer of 2001, and was followed by a workshop held in July 2001 in Toulouse, in conjunction with ACL 2001. Senseval-2 included tasks for Basque, Chinese, Czech, Danish, Dutch, English, Estonian, Italian, Japanese, Korean, Spanish, Swedish.

Senseval-3 took place in March-April 2004, followed by a workshop held in July 2004 in Barcelona, in conjunction with ACL 2004. Senseval-3 included 14 different tasks for core word sense disambiguation, as well as identification of semantic roles, multilingual annotations, logic forms, subcategorization acquisition.

Semeval-1 / Senseval-4 is currently underway. Check the Semeval-1/Senseval-4 site for more information.

News

- The call for task proposals for Semeval-1/Senseval-4 has been issued. Task proposals are due on July 1 2006. Check the Semeval-1/Senseval-4 site for details.
- All data sets used during the Senseval-3 evaluations are now in the public domain. The Senseval-3 proceedings and panel minutes are also available. Check the Senseval-3 website.

Site maintained by Rada Mihalcea, hosted by University of North Texas

Start Dokument1 - Micros... Kalender - Microsoft... 4 Microsoft Power... Word Sense Disamb... Senseval web page ... DE 15:45

Examples of Senses of the Word “Band” from SENSEVAL

band 532732 strip n band/2/1
band 532733 stripe n band/2/1.2
band 532734 range n band/2/2
band 532735 group n band/1/2
band 532736 mus n band/1/1
band 532744 brass n brass_band
band 532745 radio n band/2/2.1
band 532746 vb v band/1/3
band 532747 silver n silver_band
band 532756 steel n steel_band
band 532765 big n big_band
band 532782 dance n dance_band
band 532790 elastic n elastic_band
band 532806 march n marching_band

band 532814 man n one-man_band
band 532838 rubber n rubber_band
band 532903 ed n band/2/3
band 532949 saw n band_saw
band 532963 course n band_course
band 532979 pl n band/2/4
band 533487 vb2 a band/2/5
band 533495 portion n band/2/1.3
band 533508 waist n waistband
band 533520 ring n band/2/1.4
band 533522 sweat n sweat_band
band 533580 wrist n wristband//1
band 533705 vb3 v band/2/6
band 533706 vb4 v band/2/7

Example 1:

The incidence of accents and rests, permuted through a regular space-time grid, becomes rhythmic in itself as it modifies, defines and enriches the grouping procedure. For example, a traditional American jazz <tag ???? '>band</> was subdivided into a front line (melodic) section, usually led by trumpet, and rhythm section, usually based on drums.

Example 1:

The incidence of accents and rests, permuted through a regular space-time grid, becomes rhythmic in itself as it modifies, defines and enriches the grouping procedure. For example, a traditional American jazz <tag "532736">band</> was subdivided into a front line (melodic) section, usually led by trumpet, and rhythm section, usually based on drums.

band 532736 mus n band/1/1

Example 2:

The headsail wardrobe currently consists of a non-overlapping working jib set on a furler, originally designed to cope with wind speeds between 10 and 35 knots plus. But Mary feels it is too small for the lower wind speeds, so she may introduce an overlapping furler for the 10 to 18 knot < ???? >band</>.

Example 2:

The headsail wardrobe currently consists of a non-overlapping working jib set on a furler, originally designed to cope with wind speeds between 10 and 35 knots plus. But Mary feels it is too small for the lower wind speeds, so she may introduce an overlapping furler for the 10 to 18 knot <tag "532734">band</>.

band 532734 range n band/2/2

Example 3:

The Moorsee Lake, on the edge of town, is ideal for swimming. rowing boats are also available for hire. Don't leave without hearing the village brass <tag ???? >band</> which plays three times a week.

Example 3:

The Moorsee Lake, on the edge of town, is ideal for swimming. rowing boats are also available for hire. Don't leave without hearing the village brass <tag "532744">band</> which plays three times a week.

band 532744 brass n brass_band

Example 4:

Here, suspended from Lewis's person, were pieces of tubing held on by rubber <ta ???? >bands</>, an old wooden peg, a bit of cork.

Example 4:

Here, suspended from Lewis's person, were pieces of tubing held on by rubber <tag "532838">bands</>, an old wooden peg, a bit of cork.

band 532838 rubber n rubber_band

Potential Applications of WSD

- Machine Translation
- Information Retrieval
- Dialogue systems
- Spelling correction
- …

How difficult is it?

- Upper bound: human performance
  - 98% correct for words like *bank* with a clear meaning
  - 65% for highly ambiguous words with overlapping meanings

How difficult is it?

Lower bound:

always pick the most likely sense

Word	# different meanings	Fraction most frequent meaning
behavior	3	96%
band	24	73%
slight	8	67%
aware	2	58%
float	28	14%

One sense per discourse, one sense per collocation

- ***(Yarowsky, 1995)'s Idea***: there are constraints between different occurrences of an ambiguous word within a corpus that can be exploited for disambiguation:
  - ***One sense per discourse***: The sense of a target word is highly consistent within any given document.
  - ***One sense per collocation***: nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order and syntactic relationship.

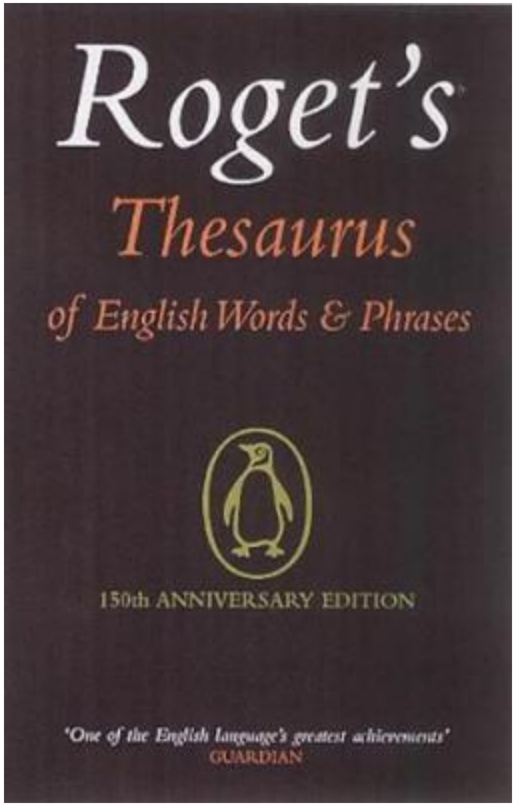
WSD Algorithms:
Data available

Methods for Disambiguating

- ***<u>Dictionary-based</u>****:* disambiguation based on lexical resources such as dictionaries and thesauri.
- ***<u>Supervised Disambiguation</u>****:* disambiguation based on a labeled training set.
- ***<u>Unsupervised Disambiguation</u>****:* disambiguation based on training on an unlabeled text corpora.

Dictionary-Based Disambiguation

- Sense definitions are extracted from existing sources such as dictionaries and thesauri.



Band - Wikipedia, the free encyclopedia - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Medien

Adresse http://en.wikipedia.org/wiki/Band Wechseln zu Links

Google wikipedia Suche 26 blockiert Rechtschreibprüfung Optionen wikipedia

Sign in / create account

article | discussion | edit this page | history

Your continued donations keep Wikipedia running!

Band

From Wikipedia, the free encyclopedia

Band may refer to:

- Band (music), a company of musicians.
- Band (radio), a range of frequencies or wavelengths between two given limits, such as those sections of the electromagnetic spectrum used in radio transmission.
- The Band, a Canadian-American rock and roll band that started in the late 1960s.
  - The Band (album), an album by The Band.
- A strip of material in a loop, used for binding, such as a rubber band.
- Bands (neckwear), a band of formal clothing fitted around the neck.
- Band society, a simple form of human society.
- The unit of First Nations government in Canada.
- Band (mathematics), an idempotent semigroup.
- a flock or herd of animals; may be sometimes used to specifically refer to gorillas. See list of collective nouns for non-human mammals.

Look up band in Wiktionary, the free dictionary.

This is a disambiguation page: a list of articles associated with the same title. If an internal link referred you to this page, you may wish to change the link to point directly to the intended article.

Categories: Disambiguation

WIKIPEDIA
The Free Encyclopedia

navigation

- Main Page
- Community Portal
- Featured articles
- Current events
- Recent changes
- Random article
- Help
- Contact Wikipedia
- Donations

search

Go Search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this article

in other languages

A WIKIMEDIA project

This page was last modified 22:35, 31 May 2006. All text is available under the terms of the GNU Free Documentation License (see Copyrights for details).
Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc.

Powered By MediaWiki

Start Dokument1 - ... Kalender - Micr... 2 Microsoft P... http://www.di... band -- Encycl... Band - Wikiped... DE 16:02

Band (music) - Wikipedia, the free encyclopedia - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Medien

Adresse http://en.wikipedia.org/wiki/Band_%28music%29 Wechseln zu Links

Google wikipedia Suche 26 blockiert Rechtschreibprüfung Optionen wikipedia

Sign in / create account

article | discussion | edit this page | history

Your continued donations keep Wikipedia running!

Band (music)

From Wikipedia, the free encyclopedia

In music, a **band** is a company of musicians, or musical ensemble, usually popular or folk, playing parts of or improvising a musical arrangement on different musical instruments.

Types of bands

[edit]

- all-women band
- big band
- brass band
- bluegrass band
- boy band
- Church band
- concert band (also known as a wind band or symphonic band)
- girl group
- jazz band
- jug band
- marching band
- military band
- pop **group**
- punk band
- rock band
- salsa band
- school band
- supergroup

WIKIPEDIA
The Free Encyclopedia

navigation

- Main Page
- Community Portal
- Featured articles
- Current events
- Recent changes
- Random article
- Help
- Contact Wikipedia
- Donations

search

Go Search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this article

in other languages

SLOX Synchronization
Appointments: Sent (2) failed item.

Start Dokument1 - ... Kalender - Micr... 2 Microsoft P... http://www.di... band -- Encycl... Band (music) - ... DE 16:03

Brass band - Wikipedia, the free encyclopedia - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten Medien
Adresse http://en.wikipedia.org/wiki/Brass_band Wechseln zu Links
Google wikipedia Suche 26 blockiert Rechtschreibprüfung Optionen wikipedia

Your continued donations keep Wikipedia running!

Brass band

From Wikipedia, the free encyclopedia


The Lochgelly Band, a Scottish colliery band, circa 1890

A **brass band** is a musical group consisting mostly of brass instruments, often with a percussion section. In some traditions other types of instruments like a clarinet or saxophones may be added, but most traditions do not accept woodwinds as part of a brass band, it would then be termed a wind band (sometimes known as a *brass and reed band*), or a military band.

While brass instruments had long been used together in various contexts, the first modern bands were developed early in the 19th century in Prussia, when all military and government bands were issued the new technology of rotary valve instruments and instructed to use standard tuning. This allowed musicians to much more easily play with other bands and for smaller bands to be combined into large bands. A separate tradition also emerged in the United Kingdom, mainly due to the importing of the new Belgian instruments, the saxhorns, invented by Adolphe Sax. These were adopted into existing British bands, and the saxhorn bands were very successful at competitions. By the early 20th century, the instrumentation had been mostly standardised, mainly by music publishers and an impressario named John Henry Iles.

Contents [hide]

Start Dokument1 - ... Kalender - Micr... 2 Microsoft P... http://www.di... band -- Encycl... Brass band - ... DE 16:16

Lesk's Algorithm

- Use dictionary descriptions to disambiguate meanings
- Set of meanings: s_1, s_2, … s_k
- Set of descriptions:

  $D(s_1)$, $D(s_2)$, … $D(s_k)$
- Words v_j in context C of ambiguous word
- Lexicon definitions of context words: $E(v_j)$

Lesk's Algorithm

Classification algorithm

$$s_{opt} = \arg\max_{s_k} \ sim\left(D(s_k), \bigcup_{v_j \in C} E(v_j) \right)$$

Possible similarity measures

$$sim(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|} \qquad sim(X,Y) = \frac{2|X \cap Y|}{|X \cup Y|}$$

$$sim(X,Y) = \frac{|X \cap Y|}{\sqrt{|X| \ |Y|}}$$

nltk_contrib.unimelb.tacohn.tagger.dictionary - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Medien

Adresse http://nltk.sourceforge.net/contrib-1.4/public/nltk_contrib.unimelb.tacohn.tagger.dictionary-module.html Wechseln zu Links

Google "Lesk's Dictionary based" Suche 26 blockiert Rechtschreibprüfung Optionen Lesk's Dictionary based

Module nltk_contrib.unimelb.tacohn.tagger.dictionary

Lesk's dictionary based disambiguator [1], and supporting dictionary interface. The dictionary interface supports both WordNet and Roget's thesaurus, as well as allowing others in the future.

This word sense tagger uses a machine readable dictionary. This simply choses the sense for each word based on the overlap between the words using in the definition of each sense in the dictionary and the words uses in the defintions of every sense for every other collocate. Collocates are simply words that appear within a window of the current word.

[1] Lesk, Michael. 1986. "Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream Cone," in Proceedings of the SIGDOC Conference

Classes	
`DictionaryI`	Abstract interface to a dictionary source, as required by the `DictionaryWordSenseTagger` class.
`LeskWordSenseTagger`	Word sense tagger using a machine readable dictionary.
`RogetDictionary`	Dictionary interface to Roget's thesaurus, as contained in the `nltk.corpus` module.
`SimulatedAnnealingWordSenseTagger`	Variant of Lesk tagger, using simulated annealing to find the optimal combination of tags for each sentence.
`WordNetDictionary`	A dictionary source using WordNet synsets as entries, using pywordnet and WordNet 1.7.

Function Summary	
	`demo()`
int	`freqdist_overlap(fdist1, fdist2)` Measures the overlap between the two distributions, in terms of the total counts for those samp

SLOX Synchronization
Appointments: Sent (2) failed item.

Start Dokument1 - Microsof... Kalender - Microsoft ... 3 Microsoft PowerP... nltk_contrib.unimelb.t... DE 09:22

Lesk's Algorithm

- Simple to implement
- No training data needed
- Relatively bad results

Supervised Disambiguation

- Training corpus:

  Each occurrence of the ambiguous word *w* is annotated with a semantic label (its contextually appropriate sense s_k).

- Supervised disambiguation is a ***classification task***.
- We will look at:
  - Bayesian classification (Gale et al. 1992).
  - Information-theoretic approach (Brown et al. 1991)

Bayesian Classification

- ***Bayes Decision rule***:

  Decide s' if

  $P(s'|C) > P(s_k|C)$ for all $s_k \neq s'$.

  (C is the context)

- Bayes decision rule is optimal because it minimizes the probability of error.
- Choose the class (or sense) with the highest conditional probability

$\mapsto$ smallest error rate.

Applying Bayes Decision Rule

- Choose a large context window around the ambiguous word.
- Combine the evidence from all features to choose the class with highest conditional probability.

Computing Posterior Probability for Bayes Classification

- Assign the ambiguous word *w* to the sense *s'*, given context *C*, where:

$$s' = \arg\max_{s_k} P(s_k|C)$$

$$s' = \arg\max_{s_k} \frac{P(C|s_k)}{P(C)} P(s_k)$$

$$s' = \arg\max_{s_k} P(C|s_k)P(s_k)$$

$$s' = \arg\max_{s_k} [\log P(C|s_k) + \log P(s_k)]$$

Naive Bayes (Gale et al. 1992)

- ***<u>Naive Bayes assumption</u>***:
  - contextual words used for description are all conditionally independent:

$$P(C \mid s_k) = \prod_{v_j \in C} P(v_j \mid s_k)$$

C: context
v_j : j-th word in C

- Consequences of this assumption:
  - ***<u>Bag of words</u>*** model: the structure and linear ordering of words within the context is ignored.
  - The presence of one word in the bag is ***<u>independent</u>*** of another.

Decision Rule for Naive Bayes

- Decide s' if

  $s' = \arg\max_{s_k} [\log P(s_k) + \Sigma_{v_j \text{ in } C} \log P(v_j | s_k)]$

- $P(v_j | s_k)$ and $P(s_k)$ are computed via Maximum-Likelihood Estimation:

$$P(v_j|s_k) = \frac{N(v_j, s_k)}{\sum_t N(v_t, s_k)} \qquad P(s_k) = \frac{N(s_k)}{N}$$

Here: N are counts

- Use your favorite smoothing technique

WSD Algorithms:

No or little or only proxy data available

WORD-SENSE DISAMBIGUATION USING STATISTICAL METHODS

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer

IBM Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

We describe a statistical technique for assigning senses to words. An instance of a word is assigned a sense by asking a question about the context in which the word appears. The question is constructed to have high mutual information with the translation of that instance in another language. When we incorporated this method of assigning senses into our statistical machine translation system, the error rate of the system decreased by thirteen percent.

http://aclweb.org/anthology-new/P/P91/P91-1034.pdf

Example of Classification based on Information-Theoretic Approach

- Two senses of a word:
  - **Prendre** une mesure <-> **take** a measure
  - **Prendre** une décision <-> **make** a decision
- The translations of the ambiguous word $\{t_1,..,t_m\}$ are {take, make, rise, speak}
- The possible indicator words $\{xf_1,\ldots,xf_n\}$ are {mesure, note, exemple, décision, parole}
- Find a partition $Q=\{Q_1, Q_2\}$ of $\{xf_1,\ldots,xf_n\}$ and $P=\{P_1, P_2\}$ of $\{t_1,\ldots,t_m\}$ that maximizes the mutual information:

$$I(P;Q)=\Sigma_{xf\in Q}\Sigma_{t\in P}p(xf,t)\ \log p(xf,t)/(p(xf)p(t))$$

Flip-Flop Algorithm (Brown et al., 1991)

- Categorize the informant (contextual word) as to which sense it indicates.

1. find a random partition $P=\{P_1, P_2\}$ for $\{t_1,\ldots,t_m\}$
2. while (improving) do
   - find partition $Q=\{Q_1, Q_2\}$ of $\{xf_1,\ldots,xf_n\}$
     - that maximizes $I(P;Q)$
   - find partition $P=\{P_1, P_2\}$ of $\{t_1,\ldots,t_m\}$
     - that maximizes $I(P;Q)$
3. end

Disambiguation using the Information-Theoretic Approach

1. For the occurrence of the ambiguous word, determine the value x_i of the indicator.
2. If x_i is in Q_1, assign the occurrence to sense 1, if x_i is in Q_2, assign the occurrence to sense 2.

Unsupervised Disambiguation

- ***<u>Idea:</u>*** disambiguate word senses without having recourse to supporting tools such as dictionaries and thesauri and in the absence of labeled text. Simply cluster the contexts of an ambiguous word into a number of groups and discriminate between these groups without labeling them.
- ***<u>(Schutze, 1998):</u>*** The probabilistic model is the same Bayesian model as the one used for supervised classification, but $P(v_j \mid s_k)$ and $P(s_k)$ are estimated using the EM algorithm.

EM algorithm

- **Initialize** the parameters μ of model. These are $P(v_j|s_k)$ and $P(s_k)$, $j = 1,2,\ldots J$, $k = 1,2,\ldots K$.
- compute the log likelihood of corpus C given the model μ: $l(C|\mu) = \log \Pi_i \Sigma_k P(c_i|s_k) P(s_k)$
- while $l(C|\mu)$ increases repeat:
  - **E-step:** $h_{ik} = P(c_i|s_k) P(s_k) / \Sigma_l P(c_i|s_l) P(s_l)$
    (use Naive Bayes to compute $P(c_i|s_k)$)
  - **M-step:** re-estimate the parameters $P(v_j|s_k)$ and $P(s_k)$ by MLE:

    $P(v_j|s_k) = \frac{\sum_i N(v_j \; in \; c_i) \cdot h_{ik}}{Z_k}$ where the sum is over all contexts c_i

    and $Z_k = \sum_j \sum_i N(v_j \; in \; c_i) \cdot h_{ik}$ is a normalizing constant.

    $P(s_k) = \Sigma_i h_{ik} / Z'_k$ with $Z'_k = \Sigma_k \Sigma_i h_{ik}$

Disambiguation

- Once the model parameters have been estimated, a word *w* can be disambiguated by computing the probability of each sense given the words v_j in the context.
- Again we use the Naïve Bayes assumption:

Decide $s' = \text{argmax}_{s_k} [\log P(s_k) + \sum_{v_j \text{ in } C} \log P(v_j | s_k)]$

Performance of Unsupervised Disambiguation

- Is capable of identifying minute difference in senses, e.g. a bank in physical sense and in abstract sense.
- Usually the clusters obtained are not identical with dictionary senses.
- Results of unsupervised disambiguation (Schütze 1998)

word	sense	Mean accuracy
suit	lawsuit	95
	garment	96
motion	physical movement	85
	proposal for action	88
train	Line of railroad cars	79
	teach	55

Learning Problems

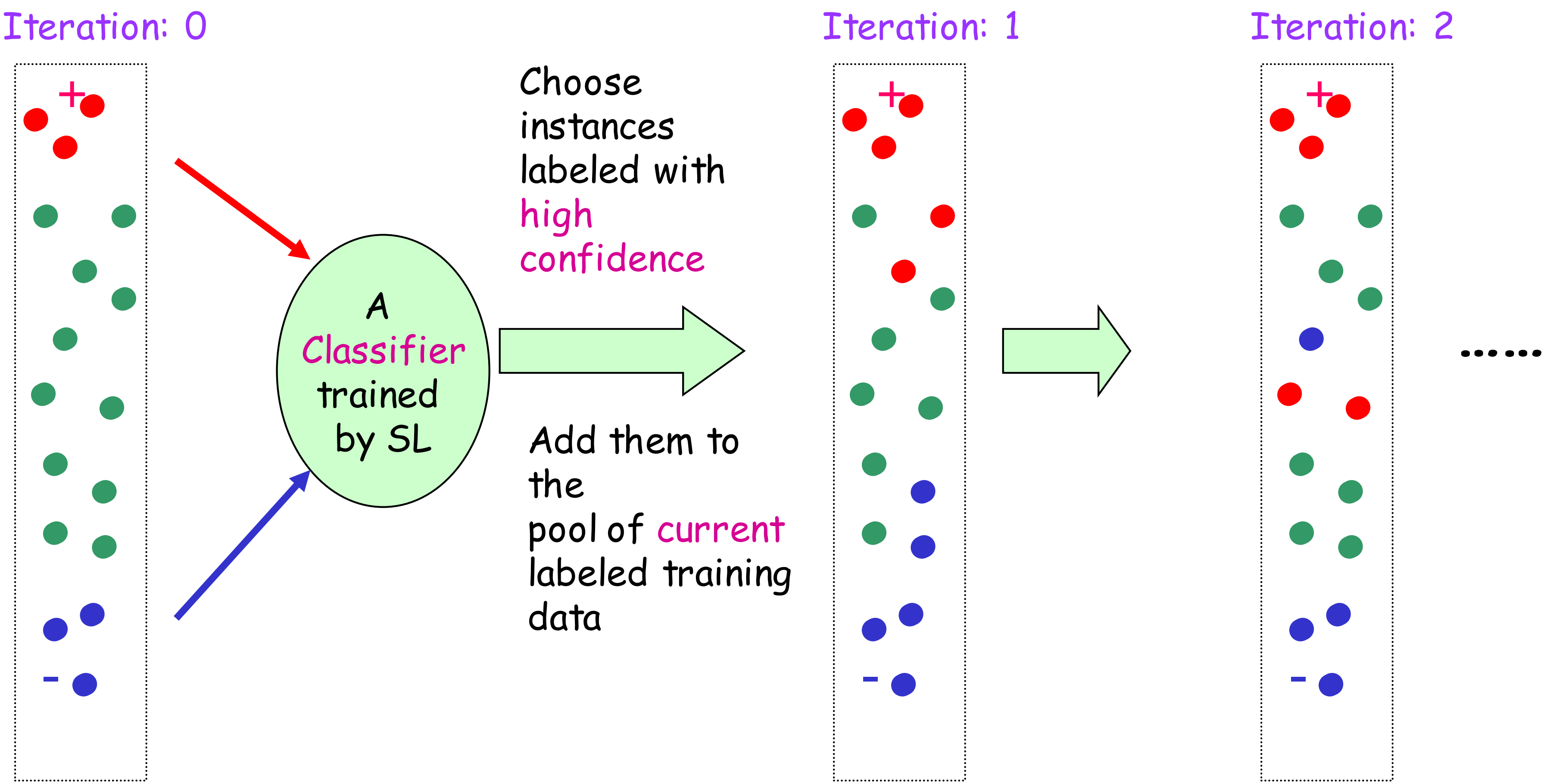
- Supervised learning:
  - Given a sample consisting of object-label pairs (x_i, y_i), find the predictive relationship between objects and labels.

- Un-supervised learning:
  - Given a sample consisting of only objects, look for interesting structures in the data, and group similar objects.

- What is Semi-supervised learning?
  - Supervised learning + additional unlabeled data
  - Unsupervised learning + additional labeled data

The Yarowsky Algorithm

(Yarowsky 1995)



Results from Yarowsky 95

UNSUPERVISED WORD SENSE DISAMBIGUATION
RIVALING SUPERVISED METHODS

David Yarowsky
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
yarowsky@unagi.cis.upenn.edu

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
					Seed Training Options			(7) + OSPD		
Word	Senses	Samp. Size	% Major Sense	Supvsd Algrtm	Two Words	Dict. Defn.	Top Colls.	End only	Each Iter.	Schütze Algrthm
plant	living/factory	7538	53.1	97.7	97.1	97.3	97.6	98.3	98.6	92
space	volume/outer	5745	50.7	93.9	89.1	92.3	93.5	93.3	93.6	90
tank	vehicle/container	11420	58.2	97.1	94.2	94.6	95.8	96.1	96.5	95
motion	legal/physical	11968	57.5	98.0	93.5	97.4	97.4	97.8	97.9	92
bass	fish/music	1859	56.1	97.8	96.6	97.2	97.7	98.5	98.8	–
palm	tree/hand	1572	74.9	96.5	93.9	94.7	95.8	95.5	95.9	–
poach	steal/boil	585	84.6	97.1	96.6	97.2	97.7	98.4	98.5	–
axes	grid/tools	1344	71.8	95.5	94.0	94.3	94.7	96.8	97.0	–
duty	tax/obligation	1280	50.0	93.7	90.4	92.1	93.2	93.9	94.1	–
drug	medicine/narcotic	1380	50.0	93.0	90.4	91.4	92.6	93.3	93.9	–
sake	benefit/drink	407	82.8	96.3	59.6	95.8	96.1	96.1	97.5	–
crane	bird/machine	2145	78.0	96.6	92.3	93.6	94.2	95.4	95.5	–
AVG		3936	63.9	96.1	90.6	94.8	95.5	96.1	96.5	92.2

$\mapsto$ better than unsupervised algorithm by Schütze

Summary

- Determine meaning of a word
- Approaches:
  - Thesaurus based
  - Supervised ($\mapsto$ text classification)
  - Unsupervised
  - Semi supervised (Yarowsky)