Computational Linguistics

Applications of Large Parsed Corpora

Gosse Bouma

Information Science
University of Groningen

LOT 2009

Overview

Word Meaning and Corpora

Meaning and Usage

You shall know a word by the company it keeps – R.Firth (1957)

- Words don't occur randomly in text
- Meaning of words determines in which texts they will be used
- Conversely: If you know in which texts and contexts a word is used (frequently), you can learn the meaning of the word.

Word Meaning and Usage

What the beep is a cobza?

The Cobza has become less widespread in usage than previously
A good cobza functions like a drum machine
The problem is, there really aren't many good cobza players left The cobza is an instrument that is dying out fast. Recorded examples of Moldavian Csango players of cobza were made during the 1950s and early 70s

Word Meaning and Usage

Cozba

The Romanian lute, known as the cobza, is a short necked, unfretted lute very similar to the oud of Iraq and Syria.

Similar Words

Applications

- Advertisements
  - Google Sponsored links
  - Which products are associated with which search terms?
- Tag Clouds and Keywords
  - What are the most important words in a text?
  - Used to give an impression of a text, blog, website
- Collocations
  - Which words do co-occur often?
  - `Geert+Wilders, hard+disk, artificial+intelligence`
  - Fixed combinations often have a special meaning, not (completely) predicatable from the meaning of their parts.

Google Ads


wielrennen - Google zoeken - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
Aanmelden
Het Internet Afbeeldingen Nieuws Maps Nieuw! Discussieg
Google
wielrennen
Zoek: het Internet pagina's in het Nederlands pagin
Het Internet Resultaten 1 - 10 van circa 4.860.000 voor wielrennen (0,11 seconden)
wielrennen.startpagina.nl
wielrennen.startpagina.nl, de beste link-pagina op het gebied van het wielrennen.
wielrennen.startpagina.nl/ - 124k -
In cache - Gelijkwaardige pagina's
Wielrensite.nl - al het wielernieuws op één site
5 13:49 - Ludovic Auger stopt met wielrennen. 3 13:45 - Schmidt wordt ploegleider bij Team CSC. 21 13:40 - Di Luca drie maanden geschorst ...
www.wielrensite.nl/ - 20k - In cache - Gelijkwaardige pagina's
Rabobank - Wielrennen
Rabobank investeert fors in wielrennen in de meest brede zin van het woord. Zo is Rabobank hoofdsponsor van de Koninklijke Nederlandse Wielren Unie (KNWU), ...
www.rabobank.nl/particulieren/sponsoring/wielrennen/ - 17k - In cache - Gelijkwaardige pagina's
Wielrennen.blog.nl
Bedankt voor je bezoek aan wielrennen.blog.nl de digitale ontmoetingsplaats voor alle aanhangers van de wielersport. ...
Gesponsorde Koppelingen
Het Laatste Sportnieuws
Nu direct het laatste sportnieuws op je iGoogle startpagina!
www.google.nl/ig
Duursportwinkel
Speciaalzaak in triathlon, duathlon zwemmen, hardlopen en fietsen
www.duursportwinkel.nl
Uw bericht hier?

Search terms for a Product

Google Ads

Search term	Advertisement
centrino	dell.nl
racefiets (*road bicycle*)	gazelle.nl
wielrennen (*cycling*)	google.nl (sportnieuws)
wielrennen (*cycling*)	vermageren.com
lance armstrong	quadrand.com (*wrist bands*)

Ad-Words and Products

Products

Companies want to find good keywords for their products

- ipod →
- hotels.nl →
- Heineken →

Finding keywords

- Is it possible to find keywords for a given product automatically?
- Which words do co-occur a lot with the name of the product?

Counting Products and Keywords

Web Counts

For a given Product and Keyword:

1. What is the number of pages containing the word Product
2. What is the number of pages containing both Product and Keyword?

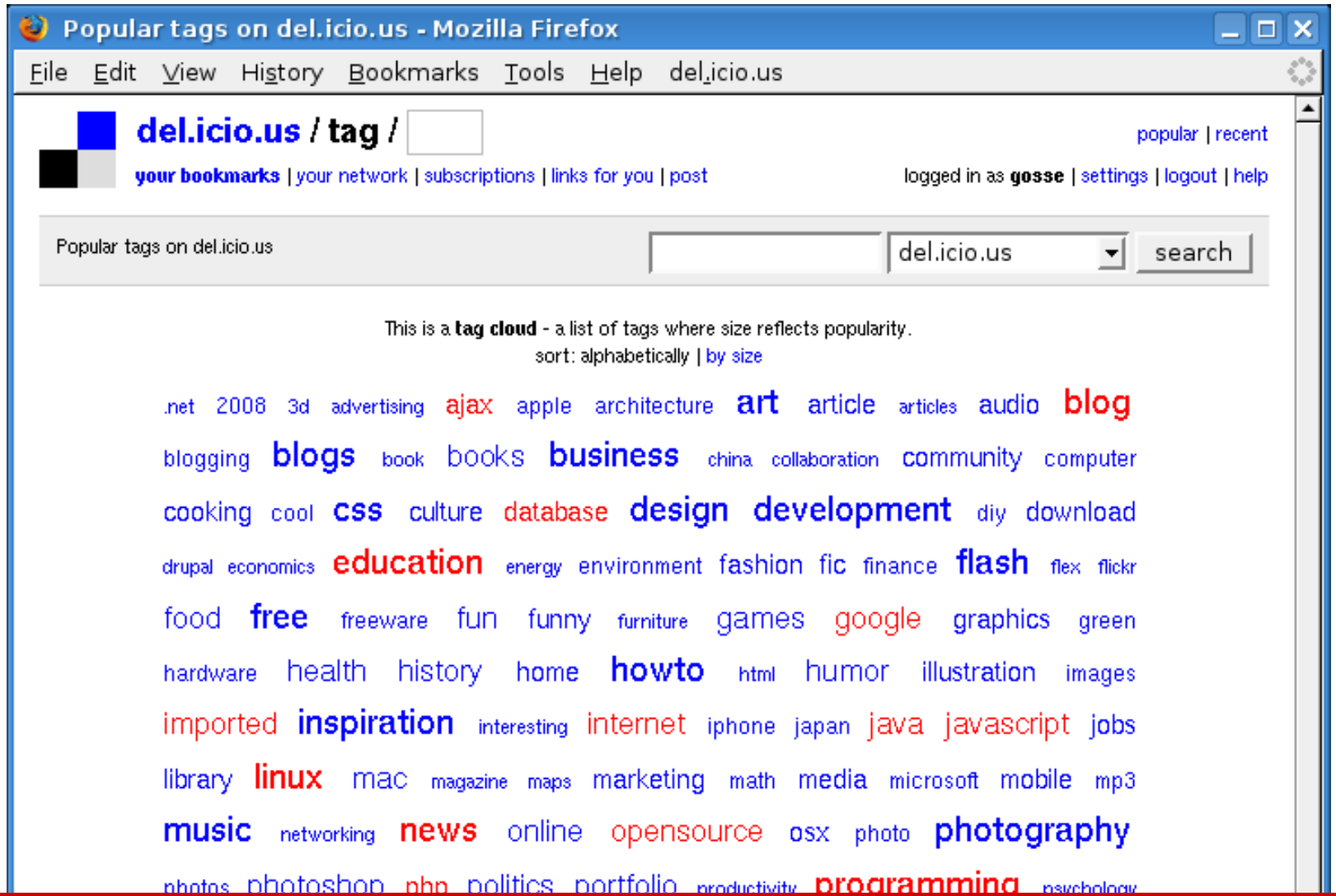
If 2 is at least 50% of 1, then Keyword might be a good keyword for Product

Tag Clouds

Finding Keywords for a Text

- Tag Clouds give an impression of the contents of a web site
  - del.icio.us
  - flickr
  - ....

del.icio.us


Popular tags on del.icio.us - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
del.icio.us / tag /
popular | recent
your bookmarks | your network | subscriptions | links for you | post
logged in as gosse | settings | logout | help
Popular tags on del.icio.us
del.icio.us
search
This is a tag cloud - a list of tags where size reflects popularity.
sort: alphabetically | by size
.net 2008 3d advertising ajax apple architecture art article articles audio blog
blogging blogs book books business china collaboration community computer
cooking cool css culture database design development diy download
drupal economics education energy environment fashion fic finance flash flex flickr
food free freeware fun funny furniture games google graphics green
hardware health history home howto html humor illustration images
imported inspiration interesting internet iphone japan java javascript jobs
library linux mac magazine maps marketing math media microsoft mobile mp3
music networking news online opensource osx photo photography
photos photoshop php politics portfolio productivity programming psychology

Keywords

From tag clouds to keywords

- Tags are keywords that are added by users
- Can we find keywords in a text (blog) automatically?
  - Good keywords are words that occur frequently in a text
- Challenges:
  - General words (and, a, the, of, by, on, is, are, have, ...) occur frequently, but are not good keywords
  - Some word-combinations (Barack Obama, open source) should be treated as a single keyword

- Make your own tag clouds: `www.tagcrowd.com`

Collocations

Fixed Expressions

- Fixed expressions are expressions of two or more words with a special, fixed, meaning
- *lingua franca, dementia praecox, habeas corpus, instant messaging, gangsta rap, ad hoc, first-person shooter, alcoholic beverages, critically acclaimed, headge fund, worth noting, mentally ill, black hole*
- *Suu Kyi, Phnom Penh, Foo Fighters, Lib Dems, Irian Jaya, Yom Kippur, Dalai Lama, ...*

Fixed Expressions

Finding fixed expressions automatically

- Fixed expressions consist of words that co-occur frequently
- However, not all frequent word-combinations are fixed expressions.
- The most frequent bigrams in a text are not (all) fixed expressions.

```
of the, in the, to the, and the, on the, by the, for the,
from the, with the, as a, of a, to be, is a, as the, is
the, at the, that the, such as, in a, ...,
```

Fixed Expressions

Finding fixed expressions automatically

- Can we do better than just most frequent combinations?
- If you read *stainless*, it is very likely that the next word is *steel*.
- For which word pairs is it the case that the combination Word1+Word2 occurs (almost) as often as Word1 or Word2 by itself?

Word Frequencies

Counts and Frequencies

- Sometimes we want to know whether word W (say, *president*) occurs more often in text A (1,000 words) than in text B (10,000 words).
- Count the number of occurrences of W in A and B ?
- (Relative) Frequency takes text size into account
  - If *president* occurs 20 times in A, the (relative) frequency of *president* in A is 20/1,000 = 0.02.
  - If *president* occurs 40 times in a text B, the (relative) frequency of *president* in B is 40/10,000 = 0.004.
- Relative Frequencies can be seen as probabilities:
  - The probability that an arbitrary word from text A is *president* is 0.02.

Fixed Expressions

What is the probability of seeing a word pair W1+W2 in a text?

Observed and Expected Frequency

- Answer 1 (Observed Frequency): the frequency of the bigram W1+W2 in the text
- Answer 2 (Expected Frequency) : the frequency of W1 multiplied with the frequency of W2.
  - As with dice: the probability of throwing $2 \times 6 = 1/6 \times 1/6 = 1/36$

Fixed Expressions vs. other bigrams

- Normal bigrams: Observed and Expected Frequency are similar
- Fixed Expressions: Observed Frequency much higher than Expected Frequency

Examples

Wikipedia fragment (approx. 10M words)

W1	W2	f(W1)	f(W2)	Observed	Expected
of	the	0.0462	0.0785	0.0147	0.0028
United	States	0.0008	0.0006	0.0005	0.000000048
stainless	steel	0.000002	0.00004	0.000002	0.0000000008

Pointwise Mutual Information

The pointwise mutual information score of a word pair W1+W2 is:

$$PMI(W1 + W2) = log \frac{f(W1 + W2)}{f(W1) \times f(W2)}$$

Wikipedia fragment (approx. 10M words)

W1	W2	f(W1)	f(W2)	Observed	PMI
of	the	0.0462	0.0785	0.0147	2.09
United	States	0.0008	0.0006	0.0005	10.26
stainless	steel	0.000002	0.00004	0.000002	14.50

Pointwise Mutual Information

Wikipedia fragment (approx. 10M words)

W1+W2	PMI	W1+W2	PMI
lingua franca	18.41	Suu Kyi	18.32
dementia praecox	17.51	Foo Fighters	18.25
habeas corpus	16.63	Mao Zedong	17.82
right-handed batsman	16.32	Alcoholics Anonymous	17.74
spinal cord	16.39	Leonhard Euler	17.57
assassination attempt	9.80	Public Library	9.32
social wellfare	9.79	Christmas Island	9.32
cable car	9.75	Cornell University	9.24
almost certainly	9.65	National Assembly	9.20
admiration for	5.74	National Council	5.58
sets out	5.74	In 1946	5.54
if they	5.74	The Simpsons	5.53
his career	5.74	The Doors	5.49

Finding Keywords

From Tags to Keywords

- Find Fixed Expressions
- Find frequent words and fixed expressions in the text,
  - But filter highly frequent words in the language in general (stopwords)

Yahoo Term Extractor

Automatic Term Extraction

Term Extraction is the task of identifying the most relevant terms in a document

- Yahoo service makes use of large (web) corpus to identify relevant terms/keywords more reliably

Yahoo Term Extractor

Vision loss may be acute or gradual; gradual vision loss is caused by multiple processes, including cataracts, glaucoma, and atrophic age-related macular degeneration. Vision loss may also be partial or complete; partial vision loss presents as visual field defects and has a variety of manifestations and causes (see Table 1: Approach to the Ophthalmologic Patient: Types of Field Defects Tab Acute vision loss may be due to central retinal artery or vein occ (including artery occlusion caused by temporal arteritis), optic r or neuropathy, vitreous hemorrhage, retinal detachment, neovascula age-related macular degeneration, stroke, or functional disorders (eg, hysterical conversion reactions or malingering).

Yahoo Term Extractor

```
Terms and wikipedia anchors in this article:

retinal artery occlusion                    -->
                http://en.wikipedia.org/wiki/retinal_artery_occlus
branch retinal artery occlusion      no wikipedia page
lit lamp examination                 no wikipedia page
vitreous hemorrhage                  no wikipedia page
closed angle glaucoma                -->
                http://en.wikipedia.org/wiki/closed_angle_glaucoma
scintillating scotoma                -->
                http://en.wikipedia.org/wiki/scintillating_scotoma
transient ischemic attacks           no wikipedia page
optic neuritis                       -->
                http://en.wikipedia.org/wiki/optic_neuritis
vein occlusion                       no wikipedia page
```

Word Meaning

Learning the Meaning of Words

- A lot of information about the meaning of a word can be learned from the contexts (sentences, texts) in which it occurs.
  - She plays the **XXX**, She studies **XXX**
  - The swinging **XXX** sounded magnificent
  - A concerto for violin, cello, trumpet, and **XXX**
- **XXX** is probably a musical instrument

Word Meaning

Word Meaning and Corpora

Corpora have been used to learn automatically

- Synonyms (two words with the same meaning: *laptop, notebook*)
- ISA-relations (*a piano is a musical instrument, a grand piano is a piano, a violin is a musical instrument*)
- Similar words: words which belong to the same category (*violin, piano, trumpet, guitar, ...*)

This information can be used to extend dictionaries automatically.

Similar Words

Contexts

Words with a similar meaning tend to occur in similar contexts What is a context?

- All the words that occur close to a word (say, at most 5 words away) in a corpus

Contexts

```
Brandenburg concerto, for solo violin, two solo flutes, strings
                  many of the violin and harpsichord concertos
          play the rapid solo violin passages.
        even arranged several violin concertos
  six sonatas and partitas for violin
  sound designer, and electric violin player
                          ...........
```

KWIC

A program which displays search results in this format is called *keyword in context* (KWIC).

Context Vector

	solo	the	of	conc erto	arra nged	elec tric	play er	sona ta
violin	100	200	50	50	10	10	30	50
piano	150	400	40	100	5	0	100	160
computer	3	600	500	3	0	300	2	0

- Which two vectors are more similar: *violin* and *piano*, or *violin* and *computer*?
- Several metrics have been proposed...

Comparing Context Vectors

	solo	the	of	concerto	arranged	electric	player	sonata
violin	100	200	50	50	10	10	30	50
piano	150	400	40	100	5	0	100	160
computer	3	600	500	3	0	300	2	0

Dice Score

$$Dice(W1, W2) = 2 \times \frac{\text{Sum of the minimum of each column}}{\text{Sum of row W1 + Sum of row W2}}$$

$$dice(viol, pia) = 2 \times \frac{100 + 200 + 40 + 50 + 5 + 0 + 30 + 50}{500 + 955} = 2 \times \frac{375}{1455} = 0.488$$

$$dice(viol, comp) = 2 \times \frac{3 + 200 + 50 + 3 + 0 + 10 + 2 + 0}{500 + 1408} = 2 \times \frac{268}{1908} = 0.280$$

Improving Context Vectors

Replacing Counts by Mutual Information

- Some context words are more informative than others
- Words like *the, of, and, is, ...* will occur frequently with most words
- Words like *sonata, concerto, ...* appear only with relatively few words
- If we fill our vectors with PMI scores instead of counts, we give more importance to words that occur relatively often with the given word.

Other Contexts

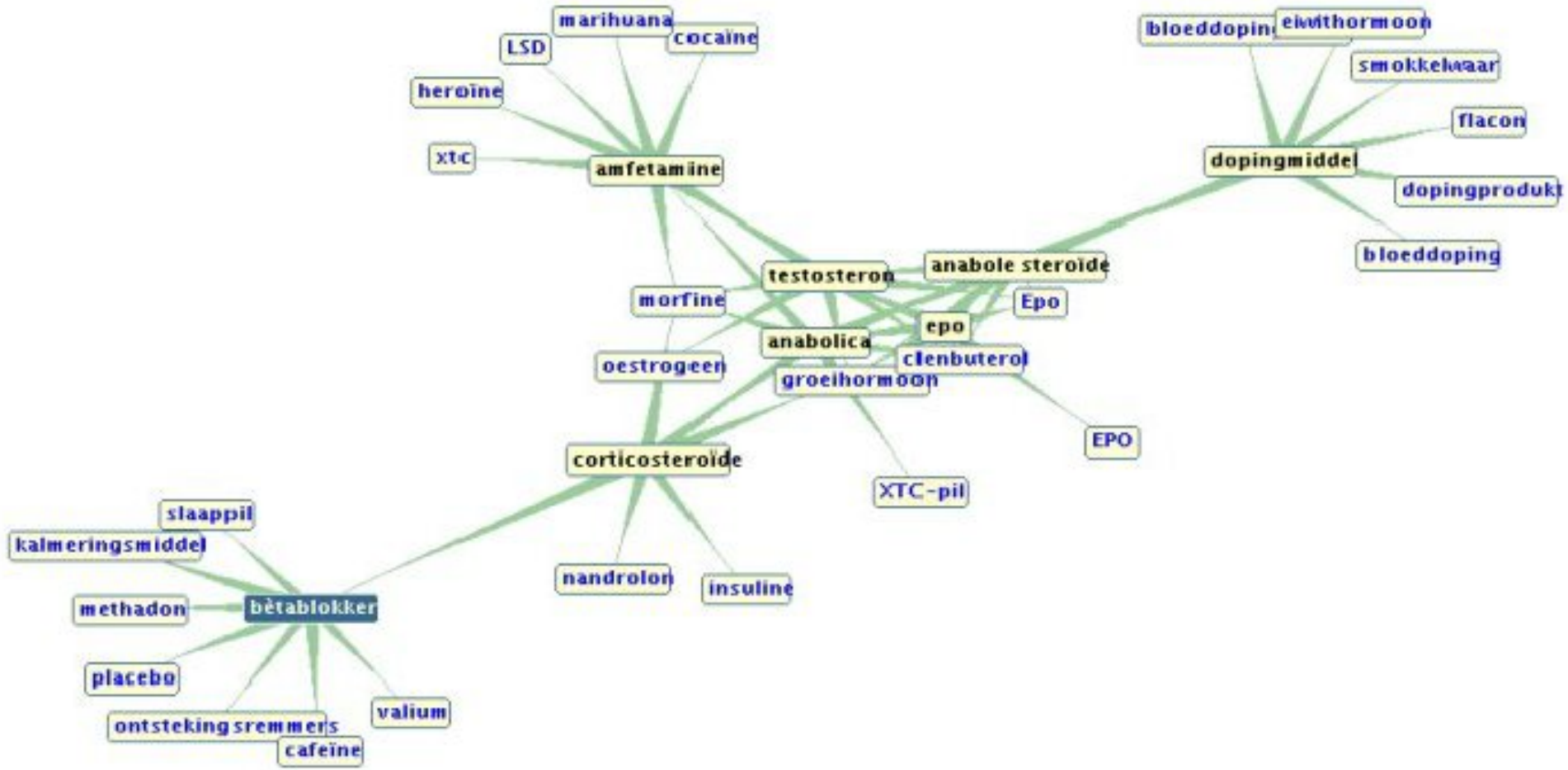
- 3 (5,10, ..) surrounding words
- All words in the sentence, in a document
- Only words in a **specific syntactic relation** to the keyword (adjectives, verbs, words in conjunctions, ...)
- Filter low frequent words (i.e. words that occur less than 5 times)

Co-occurrence data

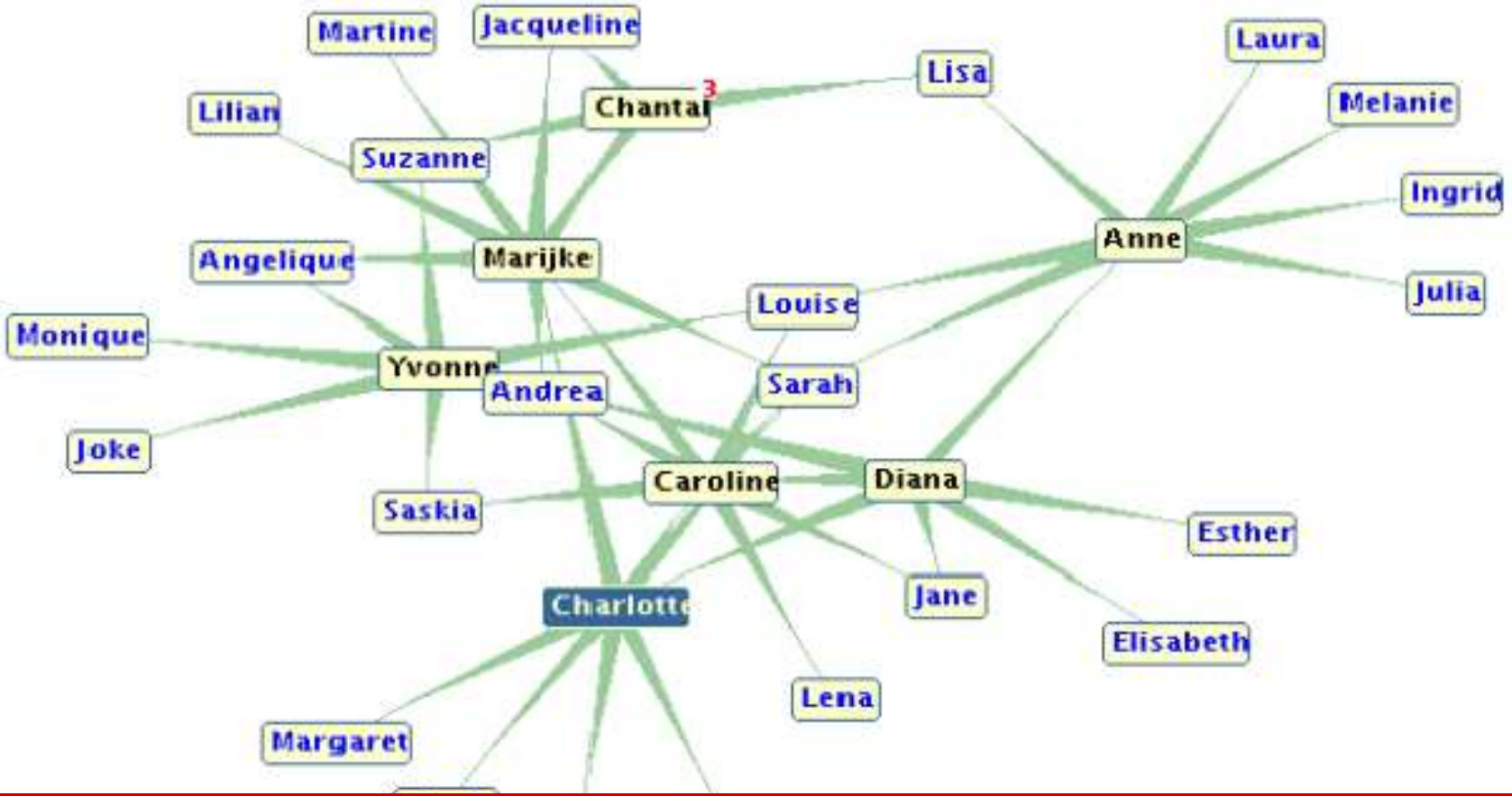
- Twente News Corpus (newspapers 1994 - 2004) (500+ M words)
- Wikipedia (25M words)

Relation	Example	Tuple	Size (M)
mod	betaalbare woning	Adj-N	21
object	woning renoveren	V-N	18
subject	woning vervuild	V-N	37
apposition	president Aristide	N-NE	12
Prep. Compl.	wisselen van woning	V+P-N	6
Coordination	woning en winkel	N-N	8

Lexical Acquisition



Lexical Acquisition



Lexical Acquisition

Noun	Apposition		Frequency
Diana	prinses	*princess*	154
Diana	vrouw	*wife*	6
Caroline	prinses	*princess*	1
Caroline	moeder	*mother*	1
Marijke	prinses	*princess*	4
Marijke	vriendin	*female friend*	3
Marijke	dochter	*daughter*	2
Marijke	moeder	*mother*	1
Yvonne	vrouw	*wife*	2
Yvonne	vriendin	*female friend*	1
Yvonne	moeder	*mother*	1
Diana	vrachtvaarder	*coaster*	2

Finding Similar Words

A lot of computation required

- Creating context vectors
  - Processing a large corpus and extracting all information
  - For Dutch syntactically annotated corpus (500M words) : 15 hrs
- Cleaning up the vector
  - Replace counts by mutual information scores
  - filter low frequency words
- Finding Similar Words
  - Dice-score between each word and all other words needs to be computed
  - For 10,000 most frequent Dutch words (nouns and names) 100,000,000 comparisons are necessary

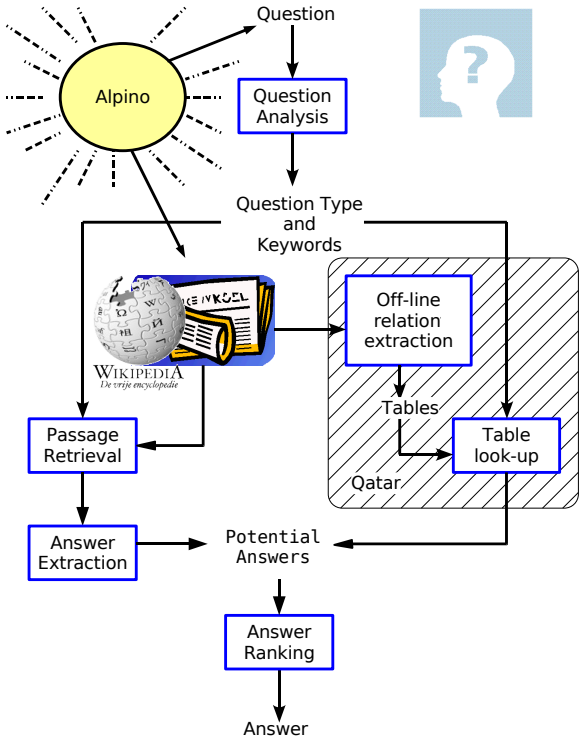
Is this useful?

- Similarity
  - Extending WordNet
  - Extending specific parts of WordNet (professions, roles of persons in organisations, ..)
- Co-occurrence data
  - Appositions used in coreference resolution and QA
- Association strength
  - Parse selection, coreference resolution

Question Answering

- When was the Ebola virus first encountered?
  - `244 persons died of the Ebola virus, that was first found in Zaire in 1976`
- How did Jimi Hendrix die?
  - `...and when on September 18, 1970, Jimi Hendrix died of an overdose, her reaction...`
- What is the capital of Russia?
  - `The riders had a tour of Moscow this morning. Tomorrow morning they are leaving the Russian capital..`

Joost: a QA system for Dutch



Dependency Relations and QA

- QA system operates on fully parsed corpora
  - CLEF (newspapers, 80M words)
  - Medical QA (reference works and web documents, 3M words)
  - On-line demo (newspapers, Wikipedia, 600M words)
- Dependency Relations used for
  - question analysis
  - answer extraction and ranking
  - information retrieval
  - off-line answer extraction
  - acquiring ontological information

Question Analysis

- Wanneer werd het Verdrag van Rome getekend?
  - *When was the Rome Treaty signed?*
  - ⟨wanneer, wh_hd, Verb⟩, ⟨Verb, su, Event⟩
  - event_date(Verdrag van Rome)
- In welke stad vond de G7 plaats?
  - *In which city did the G7 take place?*
  - ⟨in, obj, Geotype⟩, ⟨Geotype, det, welk⟩, ⟨in, wh_hd, Verb⟩, ⟨Verb, su, Event⟩
  - location(G7,stad)
- 95 patterns for 35 question types

Answer Extraction

- Given a question type,
- Search relevant documents for sentences containing a phrase that is a potential answer.

Answer Extraction

- Where did the meeting of the G7-countries take place?
  - `location(meeting,nil)`
  - `.. after a three-day meeting of the G7-countries in Napels.`
- in Napels is a potential answer for `location(meeting,nil)`
  - if NE class = LOC
  - if Answer syntactically related to Event
  - if modifiers of Event in Q and A overlap

Answer Extraction

- Answer related to Event if
  - ⟨Event, mod, Answer⟩ or
  - ⟨Verb, mod, Answer⟩, ⟨Verb, Rel, Event⟩
- Modifiers of Event_Q and Event_A overlap if
  - ⟨Event_Q, mod, Mod⟩ ↔ ⟨Event_A, mod, Mod⟩

- Where did the Olympics of 1996 take place?
- `location(Olympics)`
- `49% of the Americans does not know the Olympics of 1996 will take place in Atlanta`

Frequently Asked Question Types

- For frequenty asked question types, answers are searched off-line
  - How many inhabitants does Location have?
  - When was Person born?
  - Who won the Nobelprize for literature in 1990?
  - What does the abbreviation ADHD mean?
  - What causes Frei syndrome?
  - What are the symptoms of poisoning by mushrooms?

Relation Extraction

- Find all instances of a given relation in the corpus
  - Abbreviation & Full Term
  - Country & Capital
  - Person & Date of birth
  - Book & author
  - Movie & actor
  - Disease & Treatment
  - Term & Definition
- Search corpus using hand-crafted syntactic patterns

Organization & Founder

- ⟨ oprichten, su, Founder ⟩, ⟨ oprichten, obj, Org'n ⟩
- Minderop richtte de Tros op toen ....
- Op last van generaal De Gaulle in Londen richtte verzetsheld Jean Moulin in mei 1943 de Conseil National de la Résistance (CNR) op.
- Het Algemeen Ouderen Verbond is op 1 december opgericht door de nu 75-jarige Martin Batenburg.
- . ... toen de Generale Bank bekend maakte met de Belgische Post een "postbank" op te richten.

Cause – Effect

- Een andere **oorzaak van** glucose in de urine **is** een aangeboren nierafwijking waarbij ...
- De belangrijkste **oorzaak voor** het ontstaan van sinusitis **is** dan ook uitbreiding van neusverkoudheid naar de neusbijholten .
- De meeste reumatische aandoeningen **zijn het gevolg van** een stoornis in het afweersysteem ...
- De bewegende puntjes worden **veroorzaakt** door kleine troebelingen of deeltjes die in het heldere glasvocht zweven.

General WH-questions

- In which museum was the exhibition on Mondriaan?
  - `which(museum)`
- Which dynasty did Gengis Khan belong to?
  - `which(dynasty)`
- Name an evolution biologist.
  - `which(evolution biologist)`
- NE is a potential answer to which(Concept) if
  - NE ISA Concept

Acquiring ISA relations

- Corpus is searched exhaustively for ⟨ Concept, app, Instance ⟩
  - museum Hermitage, Madame Tussaud, National Gallery, ...(1.945)
  - bondscoach Guus Hiddink, Jorge Valdano, Louis van Gaal, ... (14K)
  - Argentinian newspaper La Nación, biologist Lilian Ramos, supermarket Disco, ... (1.861)
- 3.2M appositions extracted for 660K Named Entities

Answering definition questions

- Who is Benazir Bhutto?
  - *Prime minister of Pakistan*
- What are Koi?
  - *colored Japanese carp*
- What is Trans-Dniestr?
  - *unrecognized separatist republic inside Moldova*
- return most frequent concept label from ISA table
- expand with modifiers extracted from sentences where label was found

Question Answering

RuG NWO IMIX

Frequently Asked Questions
Best Questions
Recent Questions
Random Sample Questions

Sample Questions

Wat is het grootste eiland ter wereld?
wie is de president van Togo
Wie is Relus ter Beek?
Wat is de hoofdstad van Australië ?
Hoe hoog is de Rembrandttoren?
Welke munt gebruikt men in China
Wat is de hoofdstad van Malta
Wie is Jan de Bas?
Wie verloor de Slag bij Waterloo
Waar werd Monica Seles met een mes gestoken?

more ...

Joost Demo

ask joost

Wie is Frank van Harmelen

	Answer	Feedback	Score	found in
1.	Nederlandse artificiële- intel ligentiedeskundige	correct (1) wrong inexact	3.9362	*Volgens de Nederlandse artificiële- intel ligentiedeskundige Frank van Harmelen zijn neurale netwerken niet de weg naar superintelligente machines . [TwNC-02/ad2001/ad20011006/ad20011006_4426.xml]*

Definition Questions

- People often ask or search for definitions of terms and descriptions of persons
- Google define misses some definitions
  - *Wat is HIV?*
    - Google (nl), 'define: HIV': HIV is een virus, de volledige naam is Human Immunodeficiency Virus (menselijk immuundeficiëntievirus)...
    - Wikipedia: (lemma no. 5769, sentence no. 1)
  - *Wat is Röntgenkristallografie?*
    - Google (nl), 'define: röntgenkristallografie': no result.
    - Wikipedia: Röntgenkristallografie is de belangrijkste methode om de moleculaire structuur van eiwitten en andere biopolymeren te bepalen. (lemma no. 14573, sentence no. 35)

Syntactic Patterns for Definitions

- ok Een spanningspneumothorax is een ernstige en potentieel levensbedreigende vorm van pneumothorax.
- ok Een epileptische aanval is de reactie op een abnormale elektrische ontlading in de hersenen.
- ?? Goede bronnen voor deze vitamine zijn gist, varkensvlees, peulvruchten .. .
- ?? Vreemd genoeg zijn sommige jongens met fragiele-X-syndroom geestelijk normaal terwijl sommige meisjes..

Syntactic Patterns for Definitions

- Using syntactic properties of definitions.
- Extract sentences containing:
  - A nominal subject, a form of the copula zijn, a predicative complement (pattern: X is Y; Y is X).
- Includes:
  - (def) Zeolieten zijn mineralen met een poreuze structuur (X is Y)
  - (def) De derde waterstof isotoop is tritium (Y is X)
  - (nondef) Het enige zuur waarin platina oplost is koningswater
- Excludes: zijn used as auxiliary, as possessive pronoun, ...

Training Corpus

Sent Pos	Def	Non-Def	Undecided
First	831	18	31
Other	535	915	170

- Feature Selection
  - Text: words, bigrams, roots
  - Sentence Position: 1st, 2nd, 3rd, 4th, 5th, other
  - Syntactic: subject position (initial, non-initial), def/indef/other, LOC/PER/ORG/NONEC

Learning to Identify Definitions

- Trained a (MaxEnt) classifier which distinguishes definitions from non-definitions

Method	Precision
Baseline	59.4
Sentence position	75.9
MaxEnt	92.2

Syntactic Variation

Q and A sentence often have a slightly different syntactic structure

- How many nature reserves does Costa Rica have?
- `It is one of the 39 nature reserves of Costa Rica`
- Q: ⟨has, su, Costa Rica⟩, ⟨has, obj, reserves⟩,
- A: ⟨reserves, mod, of⟩, ⟨of, obj, Costa Rica⟩

Inference Rules over Dep Rels

- Passive
  - Asylum was given to Mengistu by Zimbabwe
    → `Zimbabwe gave asylum`
  - ⟨be,vc,Verb⟩, ⟨Verb,mod,by⟩, ⟨by, obj, Dep⟩ → ⟨Verb,su,Dep⟩
- Coordination
  - Jeltsin and Clinton will attend the G7
    → `Jeltsin will attend the G7`
  - ⟨Head, Rel, Crd⟩, ⟨Crd, cnj, Dep⟩, ⟨Head, Rel, Dep⟩
    → ⟨Head,Rel,Dep⟩

Syntactic Similarity Score

{Dependency Relations}

$\Downarrow$

Inference Rules

$\Downarrow$

{Dependency Relations & Inferred Dep Rels}

- Syntactic Similarity
  - Proportion of dependency relations R from Q for which a lexically equivalent dependency relation R' can be found in A

CLEF results

	# Q	# ok	% ok
Factoid Questions	114	62	54.4
Temporally Restricted Questions	26	7	26.9
Definition Questions	60	30	50.0
`gron051nlnl` (3 rd)	200	99	49.5
CLEF QA Monolingual Avg (42 runs)			29.5
CLEF QA Monolingual Best	200	129	64.5

Errors: Phrasal Projection ≠ Answer

- What is Hubble?
  - the repaired space telescope Hubble
- What is OJ Simpson accused of?
  - who is accused of murder on his ex-wife
- Who controlled Sudan in 1899?
  - when *England and* Egypt controlled Sudan
- What is a cincinatto?
  - somebody *who moved to the country*

Errors: Modality and Negation

- When was the German Reunification?
  - `As early as 1962, he predicted the German...`
- Who won Wimbledon?
  - `Lendl will never win Wimbledon, says...`
- What is the height of the Eiffel Tower?
  - `The fact that the Eiffel Tower is not 18 cm high, is irrelevant to him`
- When did Suriname become independent?
  - `If it had been up to the population, Suriname would never have become independent in 1975`

Errors: Too much sensitivity to Syntax

- Who is the president of France?
  - `In July, Jean-Marie Leblanc is the president of France`
- What is the capital of Ireland?
  - `Liverpool, which is sometimes called the capital of Ireland`