Word Meaning and Corpora

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

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, hedge 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.

Examples

Wikipedia fragment (approx. 10M words)

<table>
<thead>
<tr>
<th>W1</th>
<th>W2</th>
<th>f(W1)</th>
<th>f(W2)</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>the</td>
<td>0.0462</td>
<td>0.0785</td>
<td>0.0147</td>
<td>0.0028</td>
</tr>
<tr>
<td>United</td>
<td>States</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.0005</td>
<td>0.000000048</td>
</tr>
<tr>
<td>stainless</td>
<td>steel</td>
<td>0.000002</td>
<td>0.00004</td>
<td>0.000002</td>
<td>0.000000008</td>
</tr>
</tbody>
</table>

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

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)

<table>
<thead>
<tr>
<th>W1</th>
<th>W2</th>
<th>f(W1)</th>
<th>f(W2)</th>
<th>Observed</th>
<th>PMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>the</td>
<td>0.0462</td>
<td>0.0785</td>
<td>0.0147</td>
<td>2.09</td>
</tr>
<tr>
<td>United</td>
<td>States</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.0005</td>
<td>10.26</td>
</tr>
<tr>
<td>stainless</td>
<td>steel</td>
<td>0.000002</td>
<td>0.00004</td>
<td>0.000002</td>
<td>14.50</td>
</tr>
</tbody>
</table>
Pointwise Mutual Information

Wikipedia fragment (approx. 10M words)

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

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 Tables). Acute vision loss may be due to central retinal artery or vein occlusion (including artery occlusion caused by temporal arteritis), optic or neuropathy, vitreous hemorrhage, retinal detachment, neovascular age-related macular degeneration, stroke, or functional disorders (eg, hysterical conversion reactions or malingering).
Terms and wikipedia anchors in this article:

retinal artery occlusion -->
http://en.wikipedia.org/wiki/retinal_artery_occlusion
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

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 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.
**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**

<table>
<thead>
<tr>
<th></th>
<th>solo</th>
<th>the</th>
<th>of</th>
<th>concerto</th>
<th>arranged</th>
<th>elec tric</th>
<th>play</th>
<th>sonata</th>
</tr>
</thead>
<tbody>
<tr>
<td>violin</td>
<td>100</td>
<td>200</td>
<td>50</td>
<td>50</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>piano</td>
<td>150</td>
<td>400</td>
<td>40</td>
<td>100</td>
<td>5</td>
<td>0</td>
<td>100</td>
<td>160</td>
</tr>
<tr>
<td>computer</td>
<td>3</td>
<td>600</td>
<td>500</td>
<td>3</td>
<td>0</td>
<td>300</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

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

**Comparing Context Vectors**

<table>
<thead>
<tr>
<th></th>
<th>solo</th>
<th>the</th>
<th>of</th>
<th>concerto</th>
<th>arranged</th>
<th>elec tric</th>
<th>play</th>
<th>sonata</th>
</tr>
</thead>
<tbody>
<tr>
<td>violin</td>
<td>100</td>
<td>200</td>
<td>50</td>
<td>50</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>piano</td>
<td>150</td>
<td>400</td>
<td>40</td>
<td>100</td>
<td>5</td>
<td>0</td>
<td>100</td>
<td>160</td>
</tr>
<tr>
<td>computer</td>
<td>3</td>
<td>600</td>
<td>500</td>
<td>3</td>
<td>0</td>
<td>300</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

**Dice Score**

\[
\text{Dice}(W_1, W_2) = 2 \times \frac{\text{Sum of the minimum of each column}}{\text{Sum of row } W_1 + \text{Sum of row } W_2}
\]

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

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

**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)

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

Lexical Acquisition

<table>
<thead>
<tr>
<th>Noun</th>
<th>Apposition</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diana</td>
<td>prinses</td>
<td>princess</td>
</tr>
<tr>
<td>Diana</td>
<td>vrouw</td>
<td>wife</td>
</tr>
<tr>
<td>Caroline</td>
<td>prinses</td>
<td>princess</td>
</tr>
<tr>
<td>Caroline</td>
<td>moeder</td>
<td>mother</td>
</tr>
<tr>
<td>Marijke</td>
<td>prinses</td>
<td>princess</td>
</tr>
<tr>
<td>Marijke</td>
<td>vriendin</td>
<td>female friend</td>
</tr>
<tr>
<td>Marijke</td>
<td>dochter</td>
<td>daughter</td>
</tr>
<tr>
<td>Marijke</td>
<td>moeder</td>
<td>mother</td>
</tr>
<tr>
<td>Yvonne</td>
<td>vrouw</td>
<td>wife</td>
</tr>
<tr>
<td>Yvonne</td>
<td>vriendin</td>
<td>female friend</td>
</tr>
<tr>
<td>Yvonne</td>
<td>moeder</td>
<td>mother</td>
</tr>
<tr>
<td>Diana</td>
<td>vrachtvaarder</td>
<td>coaster</td>
</tr>
</tbody>
</table>
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

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..

QA at CLEF 2007

Question types

- Factoid
- Definition
- List
- Temporally restricted

Question types

- Waar wordt Ndonga gesproken?
- Wat is het Semantisch Web?
- Welke daimyo's vormden het Oostelijk leger bij de slag bij Sekigahara
- Hoeveel mobiele telefoons waren er in Nederland in gebruik in juni 1995?

- Monolingual and multilingual QA
- New in 2007 Task

- Wikipedia added to document collection
- Follow-up Questions

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 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.
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

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 frequently 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
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.

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.
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

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 potentiële levensbedreigende vorm van pneumothorax.
- ok Een epileptische aanval is de reactie op een abnormale elektrische ontlasting 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..

Question Answering
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

<table>
<thead>
<tr>
<th>Sent Pos</th>
<th>Def</th>
<th>Non-Def</th>
<th>Undecided</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>831</td>
<td>18</td>
<td>31</td>
</tr>
<tr>
<td>Other</td>
<td>535</td>
<td>915</td>
<td>170</td>
</tr>
</tbody>
</table>

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

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
  - \(\{\text{be}, \text{vc}, \text{Verb}\}, \{\text{Verb}, \text{mod}, \text{by}\}, \{\text{by}, \text{obj}, \text{Dep}\} \rightarrow \{\text{Verb}, \text{su}, \text{Dep}\}\)

- **Coordination**
  - Jeltsin and Clinton will attend the G7
    → Jeltsin will attend the G7
  - \(\{\text{Head}, \text{Rel}, \text{Crd}\}, \{\text{Crd}, \text{cnj}, \text{Dep}\}, \{\text{Head}, \text{Rel}, \text{Dep}\} \rightarrow \{\text{Head}, \text{Rel}, \text{Dep}\}\)

Syntactic Similarity Score

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

CLEF results

<table>
<thead>
<tr>
<th></th>
<th># Q</th>
<th># ok</th>
<th>% ok</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factoid Questions</td>
<td>114</td>
<td>62</td>
<td>54.4</td>
</tr>
<tr>
<td>Temporally Restricted Questions</td>
<td>26</td>
<td>7</td>
<td>26.9</td>
</tr>
<tr>
<td>Definition Questions</td>
<td>60</td>
<td>30</td>
<td>50.0</td>
</tr>
<tr>
<td>gron05lhnll1 (3 rd)</td>
<td>200</td>
<td>99</td>
<td>49.5</td>
</tr>
<tr>
<td>CLEF QA Monolingual Avg (42 runs)</td>
<td>200</td>
<td>129</td>
<td>64.5</td>
</tr>
<tr>
<td>CLEF QA Monolingual Best</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Errors: Phrasal Projection \(\neq\) 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 cincinnatto?
  - 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