Detecting Meaning with Sherlock Holmes*

The Annotated Holmes

How can we read better?

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Location: LT25

Outline

- An online edition of the Canon
- Teaching through Tagging
  — Interactive Lexical Semantics
- Sense Distributions in NTU-MC
- Word Sense Disambiguation
- Sentiment
- Where do we go from here?

An online edition of the Canon

The Adventure of the Readable Texts

For my students (for this class), I wanted texts

- faithful to the original
- easy to read on various devices
- aesthetically pleasing
- linked to the rich world of information of the Great Game

Sadly, no one site has given us what we want, so like many before us, we ended up producing our own.

Data! Data! Data!

First, I took a look at what was already out there:

Edition	Name
Gute	The Project Gutenberg HTML
ACD	Arthur Conan Doyle Encylopedia
BSW	Baker Street Wiki
SS	Short Stories
SHC	The Complete Sherlock Holmes Canon
Lit2Go	Lit2Go
Camden	Camden House: The Complete Sherlock Holmes
MoonFind	MoonFind: Searching for Sherlock

Selection Criteria

- Useful Metadata: When and where published, Author, Copyright, ...
- Annotation: links to TV and film versions, chronologies, definitions of words and Sherlockian scholarship
- Spoilers! Does the annotation reveal the villain?
- Font: Is it a nice serif font (like in the Strand Magazine)
- Resizability: can you read it on your phone
- Pictures: does it have the illustrations nicely embedded?
- Miscellaneous: does it have search?

Results

Table 1: Rating Online Texts of the Canon

Edition	Meta	Annotation	Spoilers	Font	Resize	Pics	Misc
ACD	A	A	F	C	F	C	Search
Gute	C	F	-	A	A	F	Book
BSW	A	A	F	B	F	F	Ads
SS	C	F	-	C	A	F	Ads
SHC	C	F	-	A	B	F	ToC
Lit2Go	A	F	-	C	F	F	Audio
Camden	A	F	-	C	F	A	ToC
MoonFind	F	F	-	C	F	F	Search

Every site had some good points, but no one is perfect.

Our Approach

http://compling.hss.ntu.edu.sg/canon

- Resizable, nice font with tufte-css
- Full table of contents, linked from each story
- Metadata in the margin: the date and place of publication, story number, which collection, *date of action?*
- Illustrations in the text ToDo
- Links to annotation at the end
  - Wikipedia
  - Arthur Conan Doyle Encylopedia
  - Bill Dolan's Study Guide

Annotation (new + ToDo)

- Locations link to google maps
  - What should we do for made-up places?
  - Should link locations to geonames

- Currency links to amount in today's currency
  done for SPEC

- Person + Organization should link to entry in ACD Encylopedia

- Texts are searchable with a site-specific google search
  metadata is hidden from indexing by the `<!--googleoff: all-->` command.

Annotation ToDo

- The sense level annotations of the *NTU Multilingual Corpus*, which links each open class word to its definition in Wordnet
  - We can also show translations and/or hypernyms
- Content level annotation like the Annotated Sherlock Holmes
- We have syntactic annotation (treebanks) for *The Adventure of the Speckled Band* which it would be good to display for those who are interested
- More from the Beacon Society's award winners, including information about word meaning, rhetorical devices, ...
- Links to facsimiles of the original stories

Translations and Glosses

- The Canon has been widely translated: we would like to also prepare editions of foreign languages, and consider the issue of linking the translations. E.g., *InLéctor*, who have produced a nice bilingual English-Spanish edition of *The Adventures of Sherlock Holmes*.

- With the multilingual wordnets, we can present glosses of the correct sense, in any of 34 languages (although not all will be complete)

- We can investigate how to choose the words to highlight
  - Low frequency (rare) word
  - Lower frequency sense of a word (uncommon meaning)
  - No direct translation — give definition
  - Word selected by e.g., Bill Dolan, my students, ...

  it is quite hard to do this nicely, ...

Some Technical Issues

- There is a lot of duplication in online resources (and I have only made it worse)
- How can we share information?
  - Put the code and meta-data on github

```
SPEC PUBLISHED The Strand Magazine in February 1892
SPEC ORDER 10
SPEC COLLECTION The Adventures of Sherlock Holmes
SPEC DORN http://www.beaconsociety.com/uploads/3/7/3/8/37380505/dorn_
SPEC TITLE The Adventure of the Speckled Band
SPEC IMAGE She lifted her veil 104.jpg
```

  are there standard ways to refer to lines?
- Is everyone happy to share information?
  - We have to make sure not to use data without permission

Conclusion

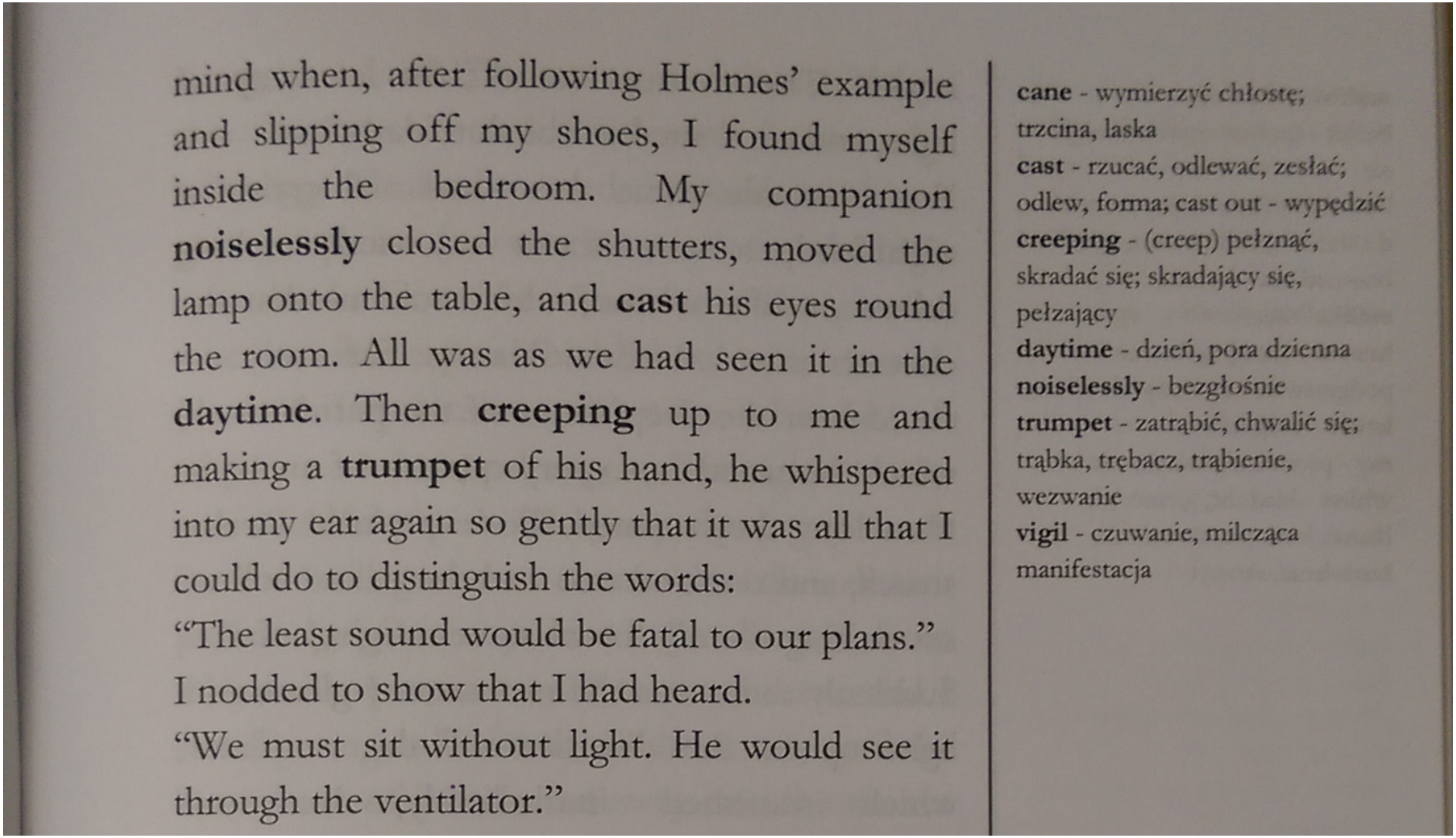
- We now have some stories with every (open-class) word linked to an entry in a dictionary, and its grammatical category
  What else can we do with this?

- For SPEC we also have this in Chinese, Japanese, Indonesian, Italian (and are working on Abui, Bulgarian, Dutch, Spanish and Polish). And these are linked to the English
  What we can do with this?

- If you have any ideas (or want the data, or want to help me make more) please let me know.

The Canon for Learners

There are some nice versions annotated for non-English speakers.



mind when, after following Holmes' example and slipping off my shoes, I found myself inside the bedroom. My companion **noiselessly** closed the shutters, moved the lamp onto the table, and **cast** his eyes round the room. All was as we had seen it in the **daytime**. Then **creeping** up to me and making a **trumpet** of his hand, he whispered into my ear again so gently that it was all that I could do to distinguish the words:

"The least sound would be fatal to our plans."

I nodded to show that I had heard.

"We must sit without light. He would see it through the ventilator."

cane - wymierzyć chłostę; trzcina, laska
cast - rzucać, odlewać, zesłać; odlew, forma; cast out - wypędzić
creeping - (creep) pełznąć, skradać się; skradający się, pełzający
daytime - dzień, pora dzienna
noiselessly - bezgłośnie
trumpet - zatrąbić, chwalić się; trąbka, trębacz, trąbienie, wezwanie
vigil - czuwanie, milcząca manifestacja

The Canon for Learners: problems

- BUT, the glosses are not very good. The correct sense is not highlighted (or sometimes even listed)
  - **cane** — whip$_v$; reed$_n$, cane$_n$
  - **cast** — throw$_v$, make forms$_v$, send out$_v$, cast ou$_v$t; metal cast$_n$, form$_n$
    should be *cast one's eyes* "take a quick look"
  - **crane** — mechanical crane
    should be *Crane Water* "place name"
  - **masonry** — "The occupation or work of a mason"
    should be *stonework*

- It is not clear how the words are selected
  Why not **dog-cart, homely, spring up, stump**?

- With our sense tags we can do better than this!

Teaching through Tagging

Teaching Meaning to my Students

- Read a Sherlock Holmes story
  - So far SPEC, DANC, REDH and SCAN
    (also news, essays and Japanese short stories)
- Look at every content word
  - Find its meaning in a dictionary (wordnet)
  - Or write a new definition if needed
  - Say if it has positive or negative connotation (for some)

  this shows how little you need to know to understand and enjoy
- Rewrite it using only the most common 1,000 words of English
  - Like XKCD's Simple Writer (used in *Thing Explainer*)
  - We only did this for REDH

  this also shows some interesting misunderstandings

Defining Meaning

- When we use a word, we don't have to know everything about the referent
  - A *dog-cart* is a kind of CART
  - $\Rightarrow$ you can ride it
  - $\Rightarrow$ it has wheels
  - $\Rightarrow$ it has something to do with a dog
- We infer that it has many of the same properties as its hypernym, even though this is not always true
  - A *hover-car* is a kind of CAR
  - $\Rightarrow$ you can ride it
  - $\not\Rightarrow$ it has wheels
- Many of the properties may be irrelevant to the story at hand, and irrelevant to the syntax of the language

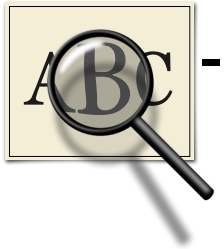
How do we learn?

You shall know a word by the company it keeps (Firth, 1957, p11)

- You see a new word in context
  buttoning up his pea-jacket,
  - ⇒ it is a kind of jacket (*green jacket*?)
  - ⇒ with buttons
  - ? it is thick material (they are going to a stake out)
  - ? it has something to do with peas
    not true (from the West Frisian word *pijjakker*, in which *pij* referred to the type of cloth used, a coarse kind of twilled blue cloth)
- And you deduce information from the context
- We are getting better at doing this with computers
  - but people use more than words: eyes, noses and other senses

How else do we learn?

- From word internal cues
  - *Television* "far vision"
  - *iphone* "internet phone" (also individual, instruct, inform, inspire)
  - 鯖 *saba* "mackerel" = 魚 fish; 青 blue
- From the sound
  - *bouba/kiki* ⋆ or ♣
  - *banged, beaten, battered, bruised, blistered, bashed*
  - mouth shape for *teeny weeny* vs *large*
- From images:



Magnifying Glass

Words are related in many other ways

- Domains: *ball, racket, net, love, ace*
- Origin: *chew, eat, drink* vs *masticate, consume, imbibe*

? come up with some words with different origins
English or another language!

- Dialect: *ripper, bonza, sickie, no worries*
- Part-of-speech: *die, live* vs *death, life*
- When you learned them!
- and many more

All of these relations affect how you use and understand language.

Idioms

- Some expressions clearly involve more than one orthographic word
  - compound noun
    - *grass snake*; *grass and tree snakes*
  - verb-particle
    - *I looked it up* vs *I looked up the very long word*
  - idiom
    - *going great guns*, *give the Devil his due*
    - *jog someone's memory*
    - *blow one's top*, *cast one's eyes*
- Knowing the individual words is not enough to know the meaning (or usage)

How common are MWEs?

- They are very common in the lexicon
  - In wordnet, 41% of the entries are multiword
- But less common in the actual text (SPEC 4.5%: 296/6,641)
  24 are new (not in Wordnet 3.0); 55 are named entities
  - *take into one's confidence*
  - *take in*
  - *Sherlock Holmes*
  - *practical joke(r)*
  - *in love*
  - *get the better of*
  - *Panama hat*
  - *as good as one's word*

Why are they important?

- If you think you know the individual words, then you might be confused
- Which is a problem if you are a translator:
  whoever crossed his path "whoever he met" SPEC
  私道を渡ろうとする人　"whoever tried to cross his private road"
- Knowledge of MWEs is one of the things that separates a good speaker from a poor one
- From a linguist's point of view, they also reveal something about how language is organized in our brains

Sense Distributions in NTU-MC

Cross-lingual comparison

We consider a single Sherlock Holmes story *The Adventure of the Speckled Band*(Conan Doyle, 1892) in the original English and translations in Mandarin Chinese, Indonesian and Japanese (NTU Multilingual Corpus: Tan and Bond, 2012). The senses are tagged with the Princeton Wordnet of English (Fellbaum, 1998), the Chinese Open Wordnet (see Wang and Bond, 2013), the Wordnet Bahasa (Bond et al., 2014) and the Japanese wordnet(Isahara et al., 2008) enhanced with pronouns, exclamatives and classifiers (COW: Seah and Bond, 2014; Morgado da Costa and Bond, 2016). On the basis of wordnet alignment within the Open Multilingual WordNet (Bond and Foster, 2013), we are able to compare the distribution of senses across languages.

Language	Sentences	Words	Concepts	MWC	SWC
English	599	11741	6425	285	6140
Indonesian	709	10345	6140	279	5861
Japanese	702	13936	4925	174	4751
Mandarin	619	12681	8263	316	7947

Table 2: Corpus size per language

		Ambiguity		Variation
Language	Lexical	Corpus	Maximum	Maximum
English	1.33	1.26	10	4
Indonesian	2.88	1.15	11	7
Japanese	1.68	1.05	9	7
Mandarin	1.30	1.01	9	10

Table 3: Ambiguity per language

- Lexical Ambiguity
  - Indonesian is high as we record both suffixed and root forms: ***igal***, ***mengigal*** "dance" (will treat as variants in OMW 2.0)
  - Japanese is high due to character variation and multiple scripts: 檜, 桧, ひのき, ヒノキ *hinoki* "Japanese cedar" (variants)
- Corpus ambiguity
  - Much closer

- ➢ Japanese and Chinese less ambiguous
- ⇒ due to the Chinese characters

One Sense Per Discourse

- Gale hypothesized that a single word would tend to have a single sense within a given discourse.
- This was definitely not the case here, in any language.
- We get from 9–11 different meanings, with (light) verbs and adverbs being most ambiguous — more true for nouns
  - en: ***see, be, so, make, have***
  - ja: **ある, 分かる, 持つ, 考える, もの**
  - zh: **好, 可能, 是, 发现, 有**
  - id: ***ada, tinggal, melihat, akhir, jadi***

Variation for a concept

- Variation (how many ways can you express the same concept) was generally lower (4-7)

- Except for Chinese, where we had 10 ways of saying "however": **还是** (1), **还** (2), **却** (11), **但** (11), **但是** (29), **然而** (2), **仍然** (1), **可是** (26), **尽管** (1), **不过** (4)
  - this is probably an artifact of the fact we have checked adverbs less, ...
  - may also be due to inconsistent tokenization

- We suspect this may change a little for different genres (future work)

Different languages are different

(1) I am sure that I shall say nothing of the kind.

a. いやいや 、そんな こと は 言わ-ん よ
iyaiya , sonna koto wa iwa-n yo
by+no+means , that+kind+of thing TOP say-NEG yo
"no no, I will not say that kind of thing"

- the components are lexicalized very differently
- *iyaiya* ↔ *I am sure that I shall* ???
- Decomposing pronouns gives us a lot of this, but the equivalence is far from direct: Can DRS help us here?

Pronomilization

(2) She$_i$ shot him$_j$ and then herself$_i$

a. 奥-さん が 旦那-さん を 撃って 、それから
oku-san ga danna-san wo utte , sorekara
wife-HON NOM husband-HON ACC shoot-CONJ , and+then
自分 も 撃った
jibun mo utta
self too shoo-PST
Wife$_i$ shot husband$_j$ and then shot self$_i$ too

➢ Why are Japanese and Chinese different here? We don't yet know, ...

Pronomilization

(3) She$_i$ shot him$_j$ and then herself$_i$

a. 她 拿 枪 先 打 丈夫 , 然后 打 自己
tā ná qiāng xiān dǎ zhàngfū , ránhòu dǎ zìjǐ
3SG take gun first shoot husband , and+then shoot self
She$_i$ took the gun to first shoot husband$_j$, and then shot self$_i$

Automatic Word Sense Disambiguation

Word Sense Disambiguation Overview

- Many words have several meanings (homonymy/polysemy)
- Determine which sense of a word is used in a specific text
- Often, the different senses of a word are closely related
  - ***title***$_1$ - right of legal ownership
  - ***title***$_2$ - document that is evidence of the legal ownership,
- sometimes, several senses can be activated in a single context
  - ...*This could bring competition to the trade*
  - ***competition***$_1$ - the act of competing
  - ***competition***$_2$ - the people who are competing

What are Word Senses?

- The meaning of a word in a given context
- Word sense representations
  - With respect to a dictionary (WordNet)
    - chair = a seat for one person, with a support for the back; "he put his coat over the back of the chair and sat down"
    - chair = the officer who presides at the meetings of an organization; "address your remarks to the chairperson"
  - With respect to the translation in a second language
    - chair = chaise
    - chair = directeur
  - With respect to the context where it occurs (discrimination)
    - "Sit on a chair" "Take a seat on this chair"
    - "The chair of the Math Department" "The chair of the meeting"

Approaches to Word Sense Disambiguation

- Knowledge-Based Disambiguation
  - Use of external lexical resources such as dictionaries and ontologies
  - Discourse properties
- Supervised Disambiguation
  - based on a labeled training set
  - basically a sequence labeling task with a lot of labels
- Unsupervised Disambiguation
  - based on unlabeled corpora
  - learn sense distinctions then disambiguate!

All Words Word Sense Disambiguation

- Attempt to disambiguate all open-class words in a text
  He put his suit over the back of the chair

- Knowledge-based approaches
  - Use information from dictionaries
  - Definitions / Examples for each meaning
  - Find similarity between definitions and current context

- Position in a semantic network
  - Find that *table* is closer to *chair* "furniture" than to *chair* "person"

- Use discourse properties
  - A word exhibits the same sense in a discourse / in a collocation

WSD with Machine Readable Dictionaries (MRD)

- MRD-based WSD shown to provide very high unsupervised baseline (e.g. Lesk algorithm in Senseval tasks)
- Suitable for all words WSD tasks (no data bottleneck)
- MRDs have (relatively) high availability compared to sensebanked data
- MRD-based WSD is easily adaptable to new MRDs, languages

What does an MRD give us?

- For each word in the language vocabulary, an MRD provides:
  - A list of meanings
  - Definitions (for all word meanings)
  - Typical usage examples (for most word meanings)
- A thesaurus adds:
  - An explicit synonymy relation between word meanings
- A semantic network/ontology adds:
  - Hypernymy/hyponymy (IS-A), meronymy/holonymy (PART-OF), antonymy, entailnment, etc.

Definitions and Examples

WordNet definitions/examples for the noun plant

1. buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles"

2. a living organism lacking the power of locomotion

3. something planted secretly for discovery by another; "the police used a plant to trick the thieves; he claimed that the evidence against him was a plant"

4. an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

Synonyms and other Relations

WordNet synsets for the noun plant

1. plant, works, industrial plant

2. plant, flora, plant life

WordNet semantic relations for the sense plant life

- hypernym: organism, being
- hyponym: house plant, fungus, ...
- meronym: plant tissue, plant part
- holonym: Plantae, kingdom Plantae, plant kingdom

Lesk Algorithm

Identify senses of words in context using definition overlap (Michael Lesk 1986)

1. Retrieve from MRD all sense definitions of the words to be disambiguated

2. Determine the **definition overlap** for all possible sense combinations

   - number of words overlapping in both definitions
   - context can be a window larger than a sentence

3. Choose senses that lead to highest overlap

Example: disambiguate *pine cone*

➢ *pine*

1. kinds of evergreen tree with needle-shaped leaves
2. waste away through sorrow or illness

➢ *cone*

1. solid body which narrows to a point
2. something of this shape whether solid or hollow
3. fruit of certain evergreen trees

$\text{pine}_1 \cap \text{cone}_1 = 0$ $\quad \text{pine}_2 \cap \text{cone}_1 = 0$

$\text{pine}_1 \cap \text{cone}_2 = 0$ $\quad \text{pine}_2 \cap \text{cone}_2 = 0$

$\text{pine}_1 \cap \text{cone}_3 = 2$ $\quad \text{pine}_2 \cap \text{cone}_3 = 0$

evergreen tree

LESK for many words

- *I saw a man who is 98 years old and can still walk and tell jokes*
- Nine open class words: see(26), man(11), year(4), old(8), can(5), still(4), walk(10), tell(8), joke(3)
- 43,929,600 sense combinations
  if we compare every definition against every definition
- How to find the optimal sense combination?
  - Find an approximate solution (e.g., simulated annealing)
  - Use a simpler algorithm

Simplified Lesk

- Original Lesk: measure overlap between sense definitions for all words in context
  - Identify simultaneously the correct senses for all words in context
  - Compare the definitions of words to the definitions of words
- Simplified Lesk: measure overlap between sense definitions of a word and current context
  - Identify the correct sense for one word at a time
  - Search space significantly reduced

Simplified Lesk Algorithm

1. Retrieve from MRD all sense definitions of the words to be disambiguated

2. Determine the overlap between each sense definition and the current context

3. Choose senses that lead to highest overlap

Disambiguate: *Pine cones hanging in a tree*

➢ PINE

1. kinds of evergreen tree with needle-shaped leaves
2. waste away through sorrow or illness

$\text{pine}_1 \cap \text{Sentence} = 1 \quad \text{pine}_2 \cap \text{Sentence} = 0$

Extended Lesk Algorithm (Banerjee and Pedersen, 2003)

1. Retrieve from MRD all sense definitions of the words to be disambiguated
   - ➢ Add definitions of hypernyms, hyponyms
   - ➢ Add definitions of the words in the definitions
2. Determine the overlap between each extended sense definition and the extended sense of each word in the context
3. Choose senses that lead to highest overlap

➢ kinds of evergreen tree with needle-shaped leaves

evergreen bearing foliage throughout the year
tree$_1$ a tall perennial woody plant having a main trunk and branches forming an elevated crown; includes gymnosperms and angiosperms
tree$_2$ tree diagram, a figure that branches from a single root; "genealogical tree"

Extended Simplified Lesk (Baldwin et al. 2009)

1. Retrieve from MRD all sense definitions of the words to be disambiguated
   - ➢ Add definitions and synonyms of hypernyms, hyponyms
   - ➢ Add definitions of the disambiguated words in the definitions
2. Determine the overlap between each extended sense definition and the each word in the context
3. Choose senses that lead to highest overlap

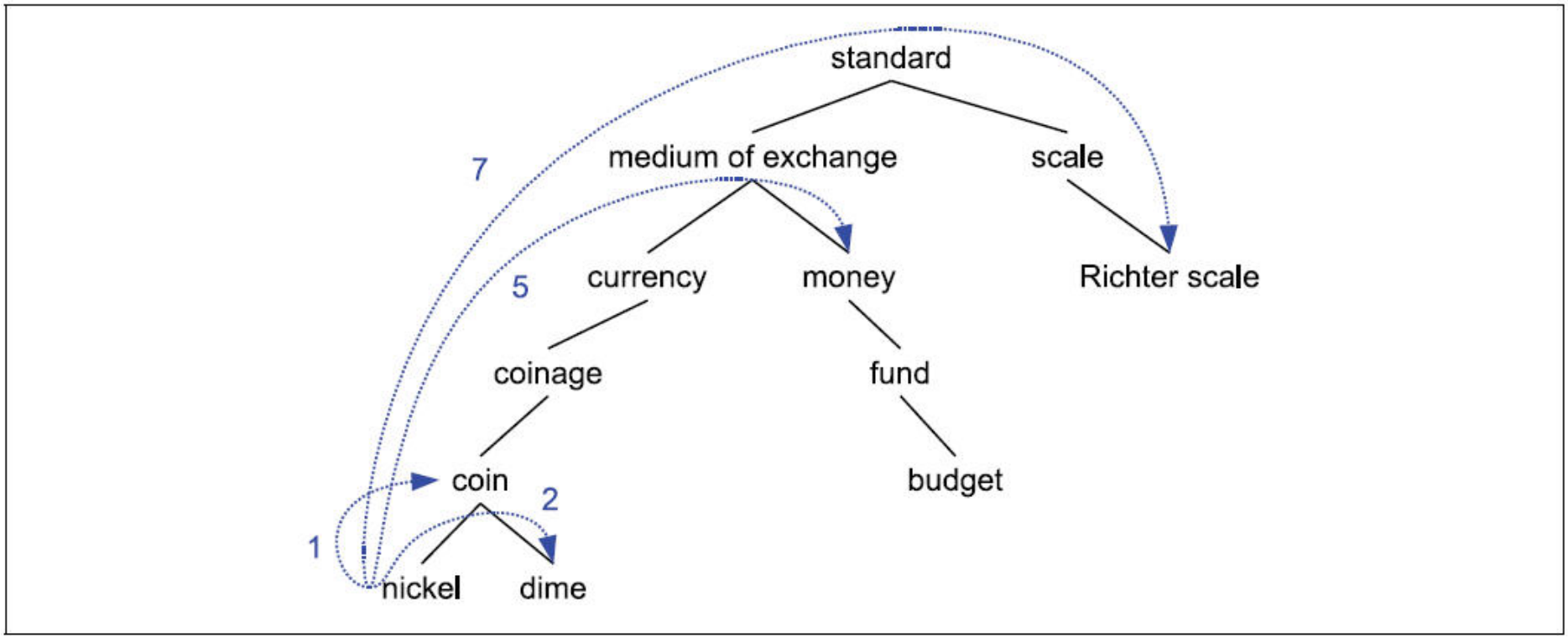
➢ kinds of evergreen$_1$ tree$_1$ with needle-shaped leaves

evergreen bearing foliage throughout the year
tree$_1$ a tall perennial woody plant having a main trunk and branches forming an elevated crown; includes gymnosperms and angiosperms

Position in a Semantic Network

- Try to find how closely related different senses are
- ...by measuring how close they are in a network
- The simplest measure is just the shortest path
  - measuring all combinations is exponential
  - normally filter by part of speech
- Better measures weight the paths
  - Small differences get low weights

Path lengths for *nickel*$_1$



➢ distance $\rightarrow$ similarity: $\text{sim}(c_1, c_2) \log \frac{1}{pathlen(c_1,c_2)}$

Corpus based Methods

- If you have a sense tagged corpus (very rare)
  - Most Frequent Sense (MFS) does very well
    - count the occurrences of each sense
    - pick the one that occurs most often
- You can improve on this with a sequence tagger, using n words of context
  - the three words on either side help (like with POS)
  - a window of 10–50 words helps!

Corpus based Learning for WSD

- Collect a set of examples that illustrate the various possible classifications or outcomes of an event.
- Identify patterns in the examples associated with each particular class of the event.
- Generalize those patterns into rules.
- Apply the rules to classify a new event.

Supervised WSD

- Learn a classifier from manually sense-tagged text using machine learning
- Resources
  - Sense Tagged Text
  - Dictionary (implicit source of sense inventory)
  - Syntactic Analysis (POS tagger, Chunker, Parser, ...)
- Scope
  - Typically one target word per context
  - Part of speech of target word resolved
  - Lends itself to some-words
- Reduces WSD to a classification problem where a target word is assigned the most appropriate sense from a given set of possibilities based on the context in which it occurs

Tagged Corpus

- Bonnie and Clyde are two really famous criminals, I think they were bank/1 robbers
- My bank/1 charges too much for an overdraft.
- I went to the bank/1 to deposit my check and get a new ATM card.
- The University of Minnesota has an East and a West Bank/2 campus right on the Mississippi River.
- My grandfather planted his pole in the bank/2 and got a great big catfish!
- The bank/2 is pretty muddy, I can' t walk there.

Bag-of-words context

bank/1 a an and are ATM Bonnie card charges check Clyde criminals deposit famous for get I much My new overdraft really robbers the they think to too two went were

bank/2 a an and big campus cant catfish East got grandfather great has his I in is Minnesota Mississippi muddy My of on planted pole pretty right River The the there University walk West

Simple Supervised Approach

- For each word w_i in S
  - If w_i is in bag-of-words(bank/1) then
    - Sense/1 = Sense/1 + 1;
  - If w_i is in bag-of-words(bank/2) then
    - Sense/2 = Sense/2 + 1;
- If Sense/1 > Sense/2 then bank/1
- else if Sense/2 > Sense/1 then bank/2
- else most frequent sense (bank/2)

Let's try it

bank/1 a an and are ATM Bonnie card charges check Clyde criminals deposit famous for get I much My new overdraft really robbers the they think to too two went were

bank/2 a an and big campus cant catfish East got grandfather great has his I in is Minnesota Mississippi muddy My of on planted pole pretty right River The the there University walk West

? I'm going to lay down my heavy load, down by the river bank.

? As a leading consumer bank in Singapore, DBS has an extensive branch and ATM network,

? My bank's Singapore headquarters is by the river at boat quay.

Commonly used features

- Identify collocational features from sense tagged data.
- Word immediately to the left or right of target: (unigram)
  - I have my bank/1 statement.
  - The river bank/2 is muddy.
- Pair of words to immediate left or right of target: (bigram)
  - The world's richest bank/1 is here in New York.
  - The river bank/2 is muddy.
- Words found within k positions around target, ($k = 10 - 50$: bag of words)
  - My credit is just horrible because my bank/1 has made several mistakes with my account and the balance is very low.

Discourse based Methods

- One sense per discourse
- One sense per collocation

One Sense per Discourse

- A word tends to preserve its meaning across all its occurrences in a discourse (Gale, Church, Yarowksy 1992)
  - 8 words with two-way ambiguity, e.g. *plant, crane, ...*
  - 98% of the two-word occurrences in the same discourse carry the same meaning
- The grain of salt: Performance depends on granularity
  - Performance of "one sense per discourse" over all words is $\approx 70\%$

One Sense per Collocation

- A word tends to preserve its meaning when used in the same collocation (Yarowsky 1993)
  - Strong for adjacent collocations
  - Weaker as the distance between words increases
- For example, in a typical corpus
  - *industrial plant* is always the plant/factory
  - *plant life* is always the plant/flora
- 97% precision on words with two-way ambiguity
- $\approx 70\%$ on all words

Typical Performance

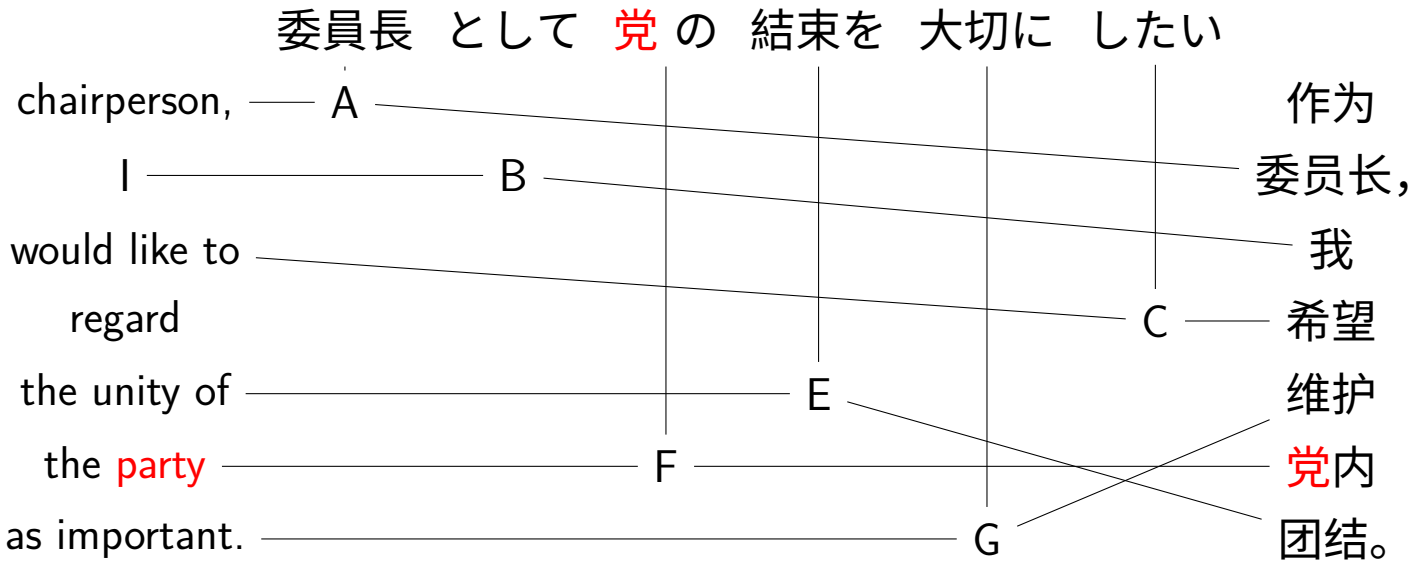
- First Sense: 63% (baseline)
- Extended Lesk: 68%
- Supervised: 70-72% (most words)
- Much harder task than POS tagging
  - Improve by reducing granularity (cluster senses)
  - Improve by increasing training data
  - Improve with more features (adding in syntax)

How can we annotate data?

- Get people to do it
  - per word (e.g. look at all *plant*) annotation much faster then per sentence
- Look at translations
  - disambiguate with other languages
- Learn collocations from unambiguous synonyms
  (*pinecone, cone, strobilus, strobile*)
- Bootstrap
  - Annotate some, assume one sense/discourse

WSD with Multiple Languages

- For multilingual corpora
  - crosslingual links narrow the interpretations
- The result is a cheaply tagged corpus



WSD with Multiple Wordnets (2)

- English
  - party$_1$ "an organization to gain political power"
  - party$_2$ "a group of people gathered together for pleasure"
  - party$_3$ "a band of people associated temporarily in some activity"
  - party$_4$ "an occasion on which people can assemble for social interaction"
- Japanese
  - 党 $_1$ "an organization to gain political power"

Summary

- There are many approaches to WSD
- We haven't solved it yet.

Sentiment

Words carry connotations

- Words can directly show how we feel about something
  - *That is good*
  - *That is awful*
- Words can indirectly show how we feel about something
  - *That is cheap*
  - *That is economical*
  - *That is old-fashioned*
  - *That is classic*
  - *That is vintage*

This is part of our knowledge of language, so it should be in the lexicon.

Some more examples

Positive	Neutral	Negative
interested	questioning	nosy
employ	use	exploit
thrifty	saving	stingy
steadfast	tenacious	stubborn
sated	filled	crammed
courageous	confident	conceited
unique	different	peculiar
meticulous	selective	picky

Can you identify the negative words?

- *East End is a gritty neighborhood, but the rents are low.*
- *On my train to Sussex, I sat next to a real stunner.*
- *Every morning my neighbor takes his mutt to the moor. It always barks loudly when leaving the building.*
- *You need to be pushy when you are looking for a job.*
- *Watson is bullheaded sometimes, but he always gets the job done.*

How can we represent this?

- One approach is a simple valence score (-100 – $+100$)

Score	Example	Example	Example	Corpus Examples
95	fantastic	very good		perfect, splendidly
64	good	good		soothing, pleasure
34	ok	sort of good	not bad	easy, interesting
0	beige	neutral		puff
-34	poorly	a bit bad		rumour, cripple
-64	bad	bad	not good	hideous, death
-95	awful	very bad		deadly, horror-stricken

- You can also have two scores (positive and negative) or even three (positive, negative and neutral)

- People also consider the associated emotion: (Plutchik, 1980)
  joy vs sadness; anger vs fear; trust vs disgust; surprise vs anticipation.

High and Low Examples

We will just look at a simple valence. Here are some results from DANC and SPEC, annotated in three languages.

Concept	freq	score	English	score	Chinese	score	Japanese	Score
i40833	24	50	marriage	39	婚事	34	結婚	58
			wedding	34				
i11080	5	40	rich	33	有钱	34	裕福	66
i72643	4	33	smile	32	微笑	34	笑み	34
i23529	40	−68	die	−80	去世	−60	亡くなる	−63
					死亡	−64	死ぬ	−62
i36562	5	−83	murder	−95	谋杀	−95	殺し	−64
							殺害	−63

Frequency is across all languages, score is average for lemma.

- Senses in the same concept tend to have similar scores
  - this is true within and across languages

Cross-Language Correlation

We looked at the agreement between annotators for the same concept across all three languages. The annotators were shown the scores organized per word and per sense: where there was a large divergence (greater than one standard deviation), they went back and checked their annotation. After this harmonization we calculated Pearson's ρ (Pearson, 1895).

Pair	ρ	**# samples**
Chinese-English	.73	6,843
Chinese-Japanese	.77	4,099
English-Japanese	.76	4,163

Bibliography

- Cheng Xiaoqing (2007) *Sherlock in Shanghai: Stories of Crime and Detection* Tr. Timothy C. Wong. Honolulu: University of Hawai' i Pres, ISBN 978-0-8248-3099-1

- Christian Metz (1974) *Film Language: A Semiotics of the Cinema* [Essais sur la signification au cinéma], Oxford University Press, 1974

- **Word Sense Disambiguation**: Jurafsky and Martin (2009), Chapter 20.1–8 figure borrowed from them

- Some slides based on Rada Mihalcea and Ted Pedersen's tutorial at AAAI-2005 "Advances in Word Sense Disambiguation"

- Nice demo of similarities at: http://maraca.d.umn.edu/cgi-bin/similarity/similarity.cgi

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL http://aclweb.org/anthology/P13-1133.

Francis Bond, Lian Tze Lim, Enya Kong Tan, and Hammam Riza. 2014. The combined wordnet Bahasa. *Nusa: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Arthur Conan Doyle. 1892. *The Adventures of Sherlock Homes*. George Newnes, London.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

J. R. Firth. 1957. *Papers in Linguistics 1934-1951*. OUP.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Luís Morgado da Costa and Francis Bond. 2016. Wow! what a useful extension to wordnet! In *10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož.

Karl Pearson. 1895. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Robert Plutchik. 1980. *Emotion: Theory, research, and experience*, volume Vol. 1. Theories of emotion. Academic, New York.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.