HG8003 Technologically Speaking:
The intersection of language and technology.

Transfer and Word Sense Disambiguation

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

Lecture 11
Location: LT8

Schedule

Lec.	Date	Topic	
1	01-16	Introduction, Organization: Overview of NLP; Main Issues	
2	01-23	Representing Language	
3	02-06	Representing Meaning	
4	02-13	Words, Lexicons and Ontologies	
5	02-20	Text Mining and Knowledge Acquisition	**Quiz**
6	02-27	Structured Text and the Semantic Web	
		Recess	
7	03-13	Citation, Reputation and PageRank	
8	03-20	Introduction to MT, Empirical NLP	
9	03-27	Analysis, Tagging, Parsing and Generation	**Quiz**
10	Video	Statistical and Example-based MT	
11	04-03	Transfer and Word Sense Disambiguation	
12	04-10	Review and Conclusions	
Exam	05-06	17:00	

- Video week 10

Introduction

- Revision:
  - EBMT
  - SMT
- Transfer in Machine Translation
- Word Sense Disambiguation

Example-based Machine Translation

Example-based Machine Translation

- When translating, reuse existing knowledge:

  0 Compile and align a database of examples
  1 Match input to a database of translation examples
  2 Identify corresponding translation fragments
  3 Recombine fragments into target text

- Example:
  - Input: He buys a book on international politics
  - Data:
    - He buys a notebook – Kare wa noto o kau
    - I read a book on international politics – Watashi wa kokusai seiji nitsuite kakareta hon o yomu
  - Output: Kare wa kokusai seiji nitsuite kakareta hon o kau

Example-based Translation: Advantages/Disadvantages

- Advantages
  - Correspondences can be found from raw data
  - Examples give well structured output if the match is big enough
- Disadvantages
  - Lack of well aligned bitexts
  - Generated text tends to be incohesive
    - Boundary Friction

Translation Memories

- **Translation Memories** are aids for human translators
  - Store and index entire existing translations
  - Before translating new text
    - Check to see if you have translated it before
    - If so, reuse the original translation

- Checks tend to be very strict $\Rightarrow$ translation is reliable
  - Identical except for white-space differences
  - The translator is in control
  - Translation companies can pool memories, giving them an advantage

Statistical Machine Translation

Statistical Machine Translation (SMT)

- Find the translation with the highest probability of being the best.
  - Probability based on existing translations (bitext)
- Balance two things:
  - Adequacy (how faithful the translation to the source)
  - Fluency (how natural is the translation)
- These are modeled by:
  - Translation Model: $P(T|S)$
    how likely is it that this translation matches the source
  - Language Model: $P(T)$
    how likely is it that this translation is good English
- Overall: $\hat{T} = \text{argmax}_T \, P(S|T) = \text{argmax}_T \, P(T|S)(T)$

Translation Model (IBM Model 4)

$$P(J, A|E)$$

Fertility Model

$$\prod n(\phi_i|E_i)$$

NULL Generation Model

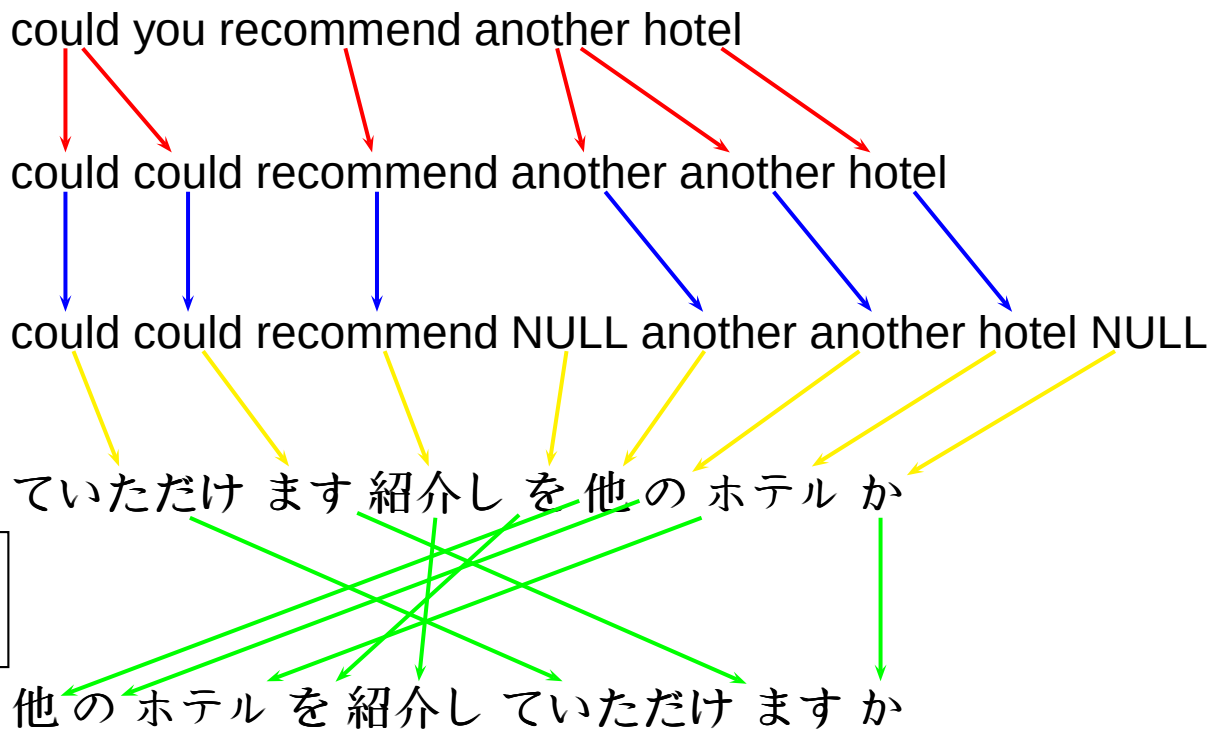
$$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$$

Lexicon Model

$$\prod t(J_j|E_{A_j})$$

Distortion Model

$$\prod d_1(j - k|\mathcal{A}(E_i)\mathcal{B}(J_j))$$
$$\prod d_{1>}(j - j'|\mathcal{B}(J_j))$$



Millions of candidates are produced and ranked.

SMT State of the Art

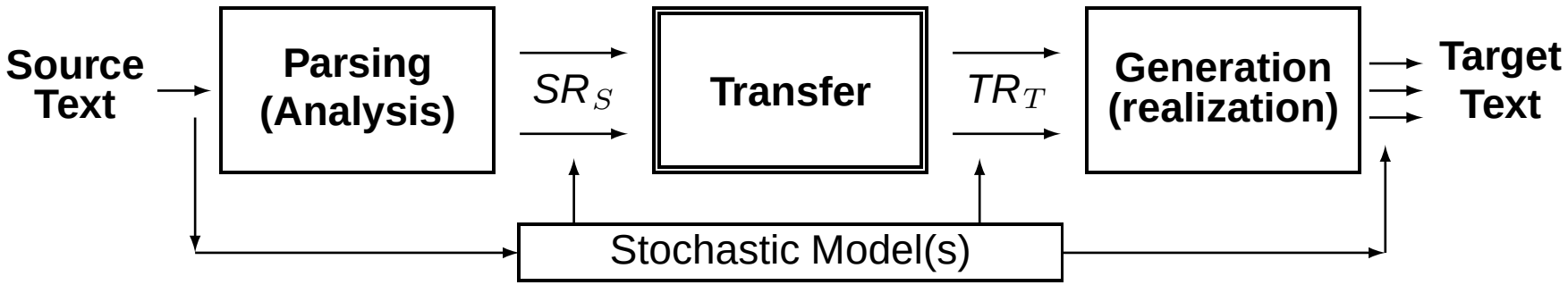
- More data improves BLEU: (Och, 2005)
  - Doubling the translation model data gives a 2.5% boost.
  - Doubling the language model data gives a 0.5% boost.
  - For linear improvement in translation quality the data must increase exponentially
    - BLEU +10% needs $2^4 = 16$ times as much bilingual data
    - BLEU +20% needs $2^8 = 256$ times as much bilingual data
    - BLEU +30% needs $2^{12} = 4096$ times as much bilingual data

Transfer

Transfer in Machine Translation

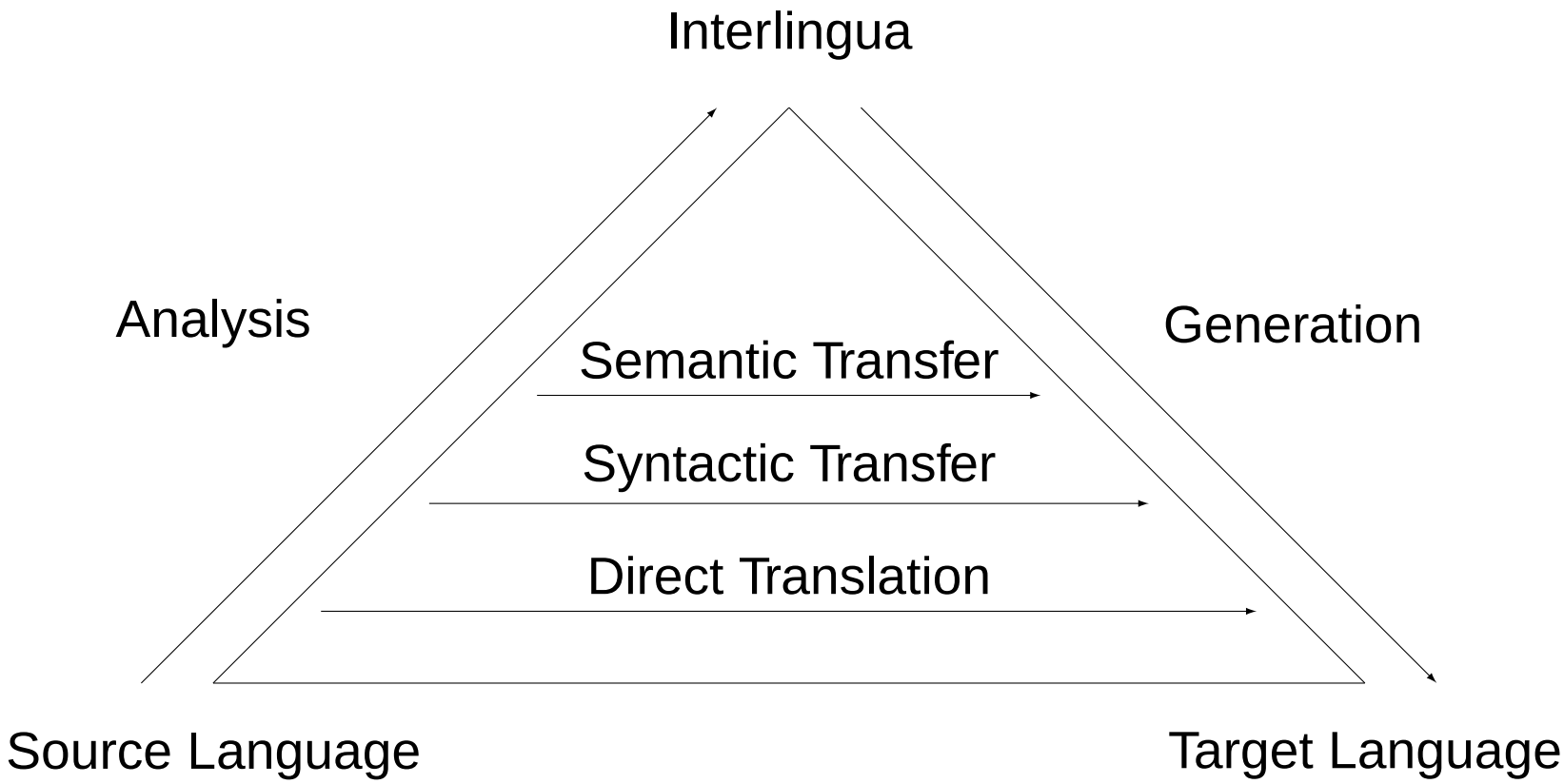
- Approaches to Transfer
- Particular Problems (and solutions)
- Ways to improve

The Overall Architecture



- Parse source text to source representation (SR)
- Transfer this to some target representation (TR) (This week)
- Generate target text from the TR

How Deep Should We Go?



The Vauquois Triangle

Direct Transfer

Input *Mary didn't slap the green witch*

Morphology Mary do-PAST NOT slap the green witch

Lexical Transfer Maria dar PAST no una bofetafa a la verde bruja

Morphology/Reordering Maria no dió una bofetafa a la bruja verde

- Just morphological analysis, no syntactic analysis
  - Works quite well for very similar languages
    - Galician/Catalan
    - Japanese/Korean
    - Malay/Indonesian
- Works very badly for languages with different word order

Lexical Selection is a problem

```
function DIRECT_TRANSLATE_MUCH/MANY(word) returns Russian translation

if preceding word is how return skol'ko
else if preceding word is as return stol'ko zhe
else if word is much
   if preceding word is very return nil
   else if following word is a noun return mnogo
else /* word is many */
   if preceding word is a preposition and following word is a noun return mnogii
   else return mnogo
```

- People write very detailed rules to select the correct translation

Japanese-English example: 鼻 *hana* “nose”

- 鼻 proper noun → Hana
- 鼻 possessed by 象 *zou* “elephant” → *trunk*
- 鼻 possessed by 馬 *uma* “horse” → *muzzle*
- 鼻 possessed by 豚 *buta* “pig” → *snout*
- 鼻 → nose
- Ontologies/thesauruses make the rules more flexible
  - mammoth ⊂ elephant
  - wild boar, hog, pig ⊂ swine
- Otherwise you have a lot of rules or miss cases

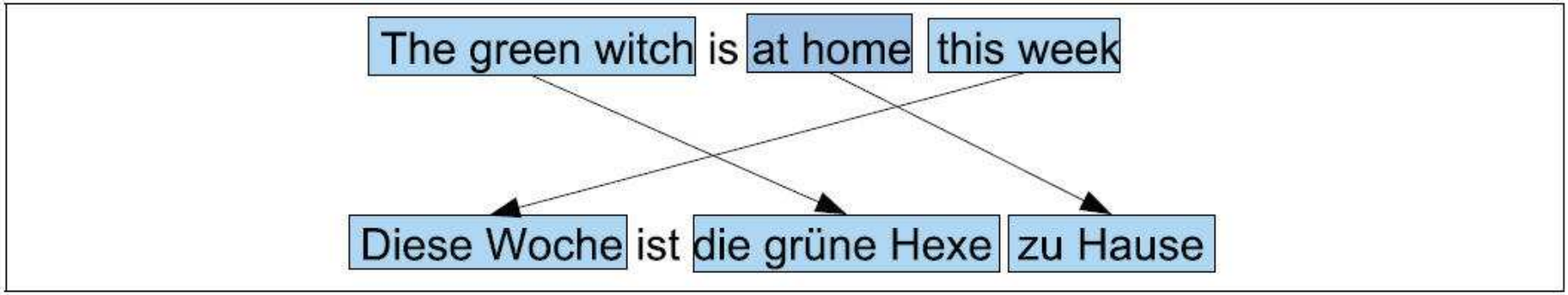
Japanese-English example: 群れ *mure* “group”

- 群れ *group* of
  - `fish` → *school* (semantic class)
  - `insect` → *swarm*
  - lion → *pride* (word)
  - wolf, wild dog → *pack*
  - star, computer → *cluster*
  - sheep → *flock*
  - `bird` → *flock*
  - `animal` → *herd*
  - `people` → *crowd*

- Many more are possible (*bevy, mob, pod, . . .*)

- This is filling in a lexical gap:
  . . . Japanese just doesn’t make these distinctions

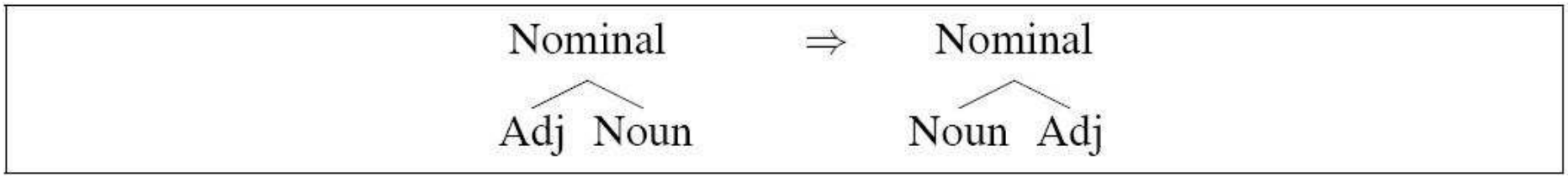
Syntactic Transfer



- Word for word won't work with very different word orders
- The condition for a transfer rule may be far away
  - *pack of wolves*
  - *pack of large, hungry, gray wolves*
- We should look at the sentence structure

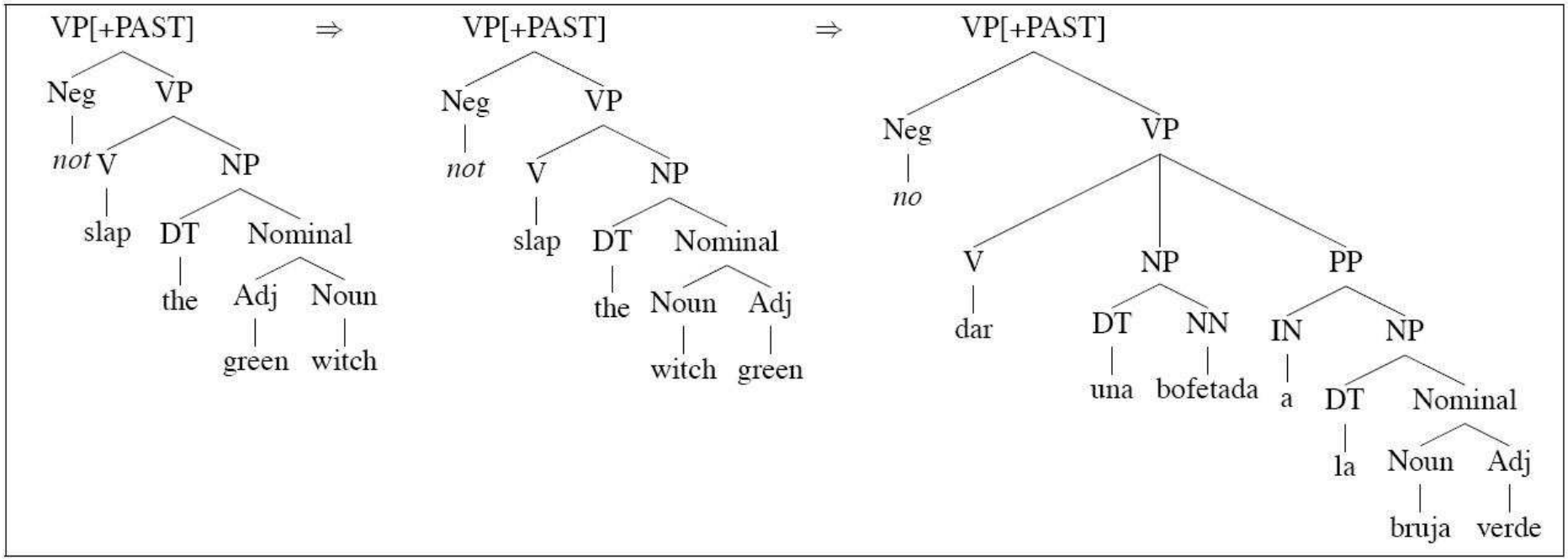
Syntactic Transfer: Spanish-English

- In Spanish, Italian, French, Malay, . . . adjectives follow nouns
  - the green witch $\rightarrow$ la bruja verde
- Try to make general rules for this

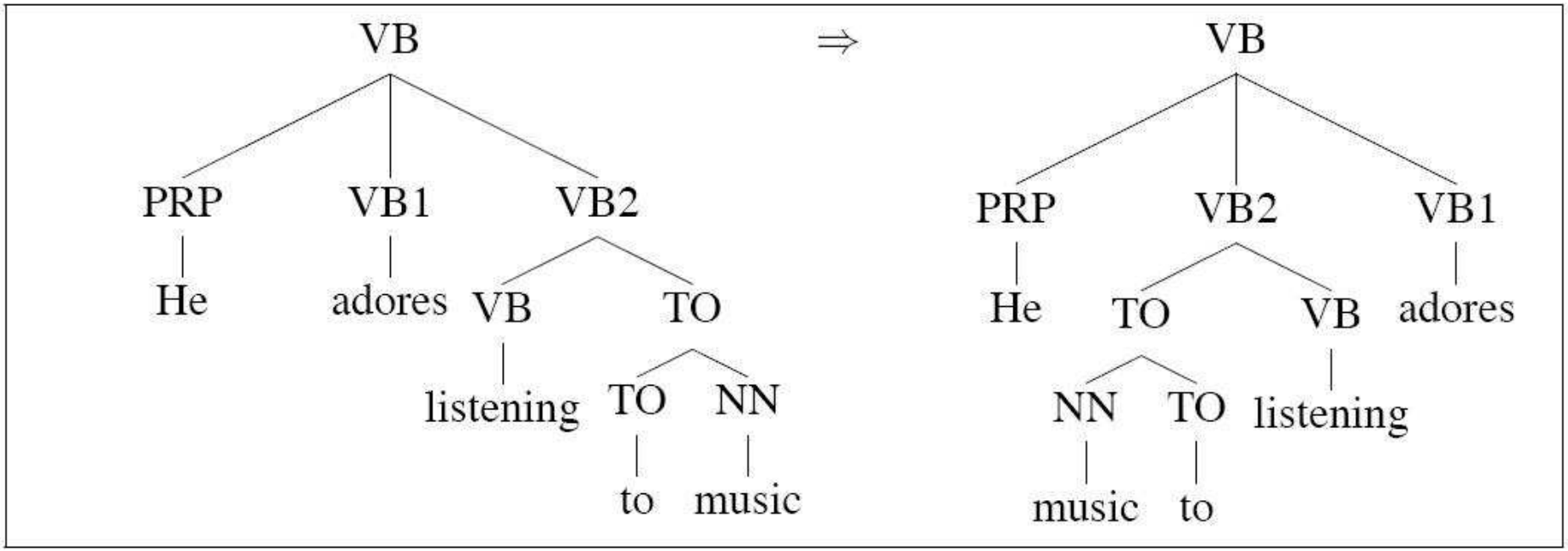


- The general strategy is to apply transfer rules top down from the root

Syntactic Transfer: Spanish-English



Syntactic Transfer: English-Japanese



- 彼が　音楽を　聞くのが　大好きだ
  kare-ga ongaku-wo kiku-no-ga daisuki-da
  he-SUBJ music-OBJ listen-NOM-SUBJ likes

- Word order is very different!

Syntactic Transfer: Various Rules

	English to Spanish:		
1.	NP → Adjective$_1$ Noun$_2$	⇒	NP → Noun$_2$ Adjective$_1$
	Chinese to English:		
2.	VP → PP[+Goal] V	⇒	VP → V PP[+Goal]
	English to Japanese:		
3.	VP → V NP	⇒	VP → NP V
4.	PP → P NP	⇒	PP → NP P
5.	NP → NP$_1$ Rel. Clause$_2$	⇒	NP → Rel. Clause$_2$ NP$_1$

Semantic Transfer

- Aim for simpler semantic transfer
  - Push work to the monolingual grammars
  - Moving toward an interlingua
  - Transfer can ignore language specific syntax
- Modularize the components
  - Define a clean **Sem**antic-**I**nterface
  - Allow independent work on components
- Reduce, Reuse, Recycle

Example: Source

- ビール を 三つ　　もって きて ください
  biiru-wo　mittsu　　motte　kite　kudasai
  beer-ACC　three-CL hold　　come give:honorific

  Please bring three beers.

- $\langle h_1, \{\mathbf{h_1}$: motsu_v($e_1$: COMMAND, u_2, $\mathbf{x_1}$),
  $\mathbf{h_1}$: kuru_v(e_2, u_3),
  h_4: biiru_n($\mathbf{x_1}$),
  h_6: udef_q($\mathbf{x_1}$, h_7, h_8),
  h_9: card(u_1, $\mathbf{x_1}$, "3"),
  h_{15}: kudasaru_v(e_3, u_4, u_5, h_2) $\}$,
  $\{h_7 = h_4, h_2 = h_1\}\rangle$

- *motte kuru* "hold come" grouped together (bring)

Example: Transfer

Transfer:

- biiru_n(x_i) $\rightarrow$ beer_n(x_i)
- h_j: motsu_v(e_1, u_2, x_1), “hold” h_j: kuru_v(e_2, u_3) “come” $\rightarrow$ h_j: bring_v(e_1, u_2, x_1)
- h_i: kudasaru_v(e_j, h_k) $\rightarrow$ h_i: please_a(e_j, h_k) (verb $\rightarrow$ adverb)

Example: Target

- $\langle h_0, \{h_0$: please_a(e_3, h_1),
  h_1: imp_m(h_3),
  h_2: pronoun_q(x_0, h_7, h_8), h_4: pron($x_0\{2nd\}$),
  h_5: bring_v($e_2, x_0, \mathbf{x_1}$),
  h_4: beer_n($\mathbf{x_1}$), h_6: udef_q($\mathbf{x_1}, h_{10}, h_8$), h_{11}: card($u_1, \mathbf{x_1}$, "3") $\}$,
  $\{h_3 = h_5, h_7 = h_4, h_{10} = h_{11}, \}\rangle$

- Two word orders possible
  - Please bring three beers.
  - Bring three beers please.

Semantic Transfer Pros and Cons

- Source and Target grammars do much of the work
  - Pro: modular, transfer easier
  - Cons: brittle (if parsing fails, everything fails)
- Language specific details hidden by the semantic interface
- General Problems Remain
  - Sense Disambiguation (lexical choice)
    is 鳩 *hato* a *dove* or a *pigeon*
  - Language Differences
    - number, countability, articles
- Over-generate and choose with a statistical model

The Importance of Multiword Expressions

- Context beyond a single word is very important
- In a typical system most rules (entries in the transfer dictionary) are multiword (60% in **ALT-J/E**)
  - 機械 翻訳 *kikai honyaku* “machine translation” → machine translation
  - 雨 が 降る *ame-ga furu* “rain falls” → rains
- If you consider conditions as part of the translation, then this goes up more
  - 鼻 *hana* “nose” possessed by 象 *zou* “elephant” → trunk
  - 鼻 *hana* “nose” possessed by 豚 *buta* “pig” → snout
  - 鼻 *hana* “nose” → nose

Issues with Transfer

- Choosing between multiple options is difficult
  - ⇒ Create larger rules with more context
  - ⇒ Try to weight with statistical models
- The number of rules is far greater than the number of words
  Context multiplies rules
  - ⇒ Generalize rules with ontologies
  - ⇒ Learn from bilingual corpora
  - ⇒ Restrict according to domain
  - ⇒ Share rules (open source)

Some well known problems

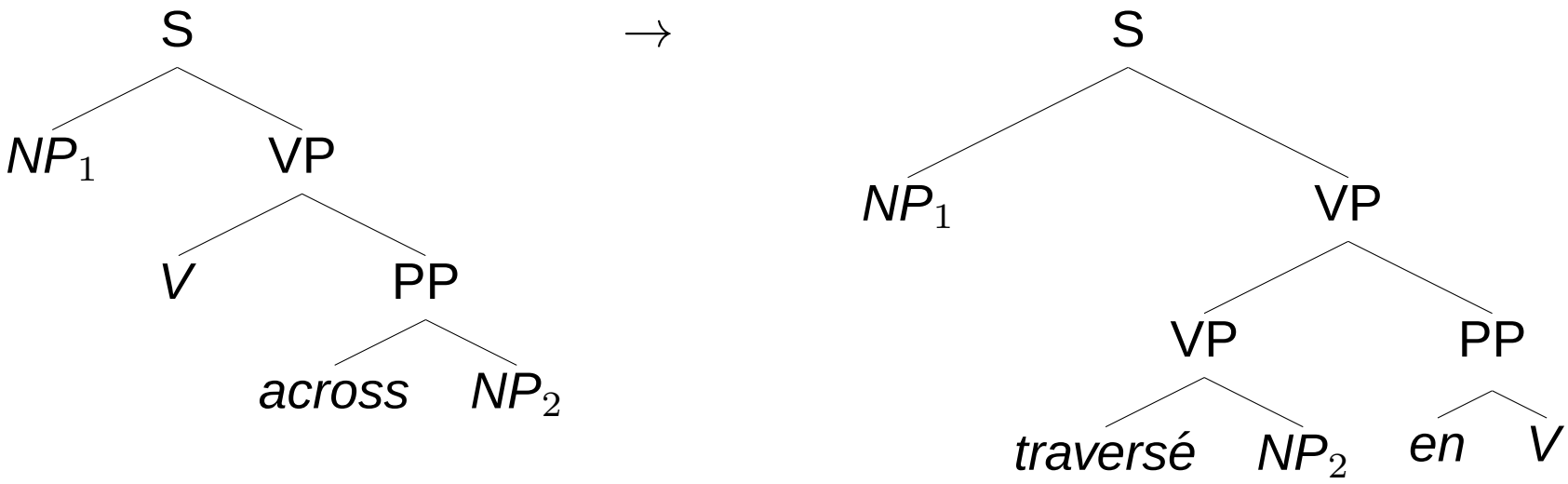
- Head-switching: head is dependent in the other language
- Relation-changing: e.g. verb $\rightarrow$ adjective
- Lexical Gaps: translation missing in the source or target language
- Possessive Pronoun Drop: possessive pronouns required in some languages, but not others
- Number mismatch: number required in one language but not the other
- Argument mismatch: Verb structure is different
- Idiom mismatch: Idiomatic in one language but not the other

Head Switching

- Head switching is just a more complicated rule:

(1) *I swam across the river*

(2) J'ai traversé le fleuve en nageant
I crossed the river by swimming



Relation Changing

- Translation equivalents may be different POS:

  (3) 濡れている 紙
  nurete iru kami
  wetting is paper
  wet paper

- Verb → Adjective

- Allow translation rules to do this
  - Normally anchor lexically to reduce complexity
    - ⊗ VP → AP
    - ∗ *nureru_v* → *wet_a*

Lexical Gaps

- More specific to less specific
  - Just lose some information
    - *herd, pack, mob, crowd, group* $\rightarrow$ *mure*
- Less specific to more specific
  - Add context to the transfer rules to disambiguate
  - Add multiword expressions to the dictionary

Possessive Pronoun Drop

REF	Kanji:	鼻が	かゆい	
	Jap:	*hana-ga*	*kayui*	
	Gloss:	nose-SUBJ	itch	
	Eng:	‘My nose itches’		
GEN	Kanji:	鼻は	感覚器官	だ
	Jap:	*hana-wa*	*kankakukikan*	*da*
	Gloss:	nose-TOP	sensory organ	is
	Eng:	‘Noses are sensory organs’		
		‘The nose is a sensory organ’		
		‘A nose is a sensory organ’		

- Possessive pronouns are obligatory for some nouns (possessed-nouns): Nouns that denote `kin, body parts, work, personal possessions, attributes` and `people defined by their relation to another person`

Generating possessive pronouns:

A If a referential phrase is headed by a possessed-noun and is not the direct object of a verb with meaning POSSESSION or ACQUISITION then:

- Generate a possessive pronoun whose referent is the subject of the sentence.

  I scratched my nose; *She scratched her nose*

B Generate possessive pronouns for all noun phrases

- Rank with a language model

- There is no perfect solution
  - **A** requires very complex processing
  - **B** makes every noun phrase very ambiguous

Number mismatch

- Some examples (Nouns are unmarked for number in Japanese)
  - マンモスは全滅した。→ *Mammoths are extinct.*
  - 花を集まった。→ *I gathered flowers.*
  - この３人は、友達だ。
    → *These three people are friends.*
  - ３人は大勢だ。→ *Three people are a crowd.*

A Write rules that use context: (accurate)

- Verb/Adjective: *be extinct*, *gather*
- Modifiers: *three*, *many*
- Defaults: *noodles*

B Over generate and rank with a language model (easy)

Argument Mismatch

- Verb (or adjective) structure is different
  - *watashi-ni kodomo-ga iru* “to me children are”
    $\rightarrow$ *I have children*
    to$\rightarrow$SUBJECT; SUBJECT$\rightarrow$OBJECT
  - *Kim married Sandy*
    $\rightarrow$ *Kim-ga Sandy-to kekkon-shita* “Kim married with Sandy”
    OBJECT$\rightarrow$*-to* “with”

Idiom mismatch

- Idiomatic in one language but not the other (or not in the same way)
  - *I lost my head* “I got angry”
    → *atama-ni kita* “it came to my head”
  - *I racked my brains* “I thought hard”
    → *chie-wo shibotta* “I squeezed knowledge”
    I lost my head → *I got angry*

- Some idioms are so common that we don’t notice them
  - *I catch the bus* “I get on the bus”
  - *I follow you* “I understand you”

User Dictionaries

- The simplest way to improve translation quality
- Build a special dictionary: the user dictionary
- User dictionary entries are preferred to words in the system dictionaries
  - You can force the translation you want
- Typical MT use for large projects is to
  1. Translate once
  2. Find common errors
  3. Fix them by adding entries to the user dictionary
  4. Re-translate

How to Predict Machine Translation Quality

- The following phenomena are hard to translate:
  - Long sentences
  - Coordination
  - Unknown words (either new words or spelling errors)
    - new genre
    - poorly edited text
  - Different language families

- We can identify these and give a translatability score
  - This is useful to identify text for post-editing

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

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

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

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

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

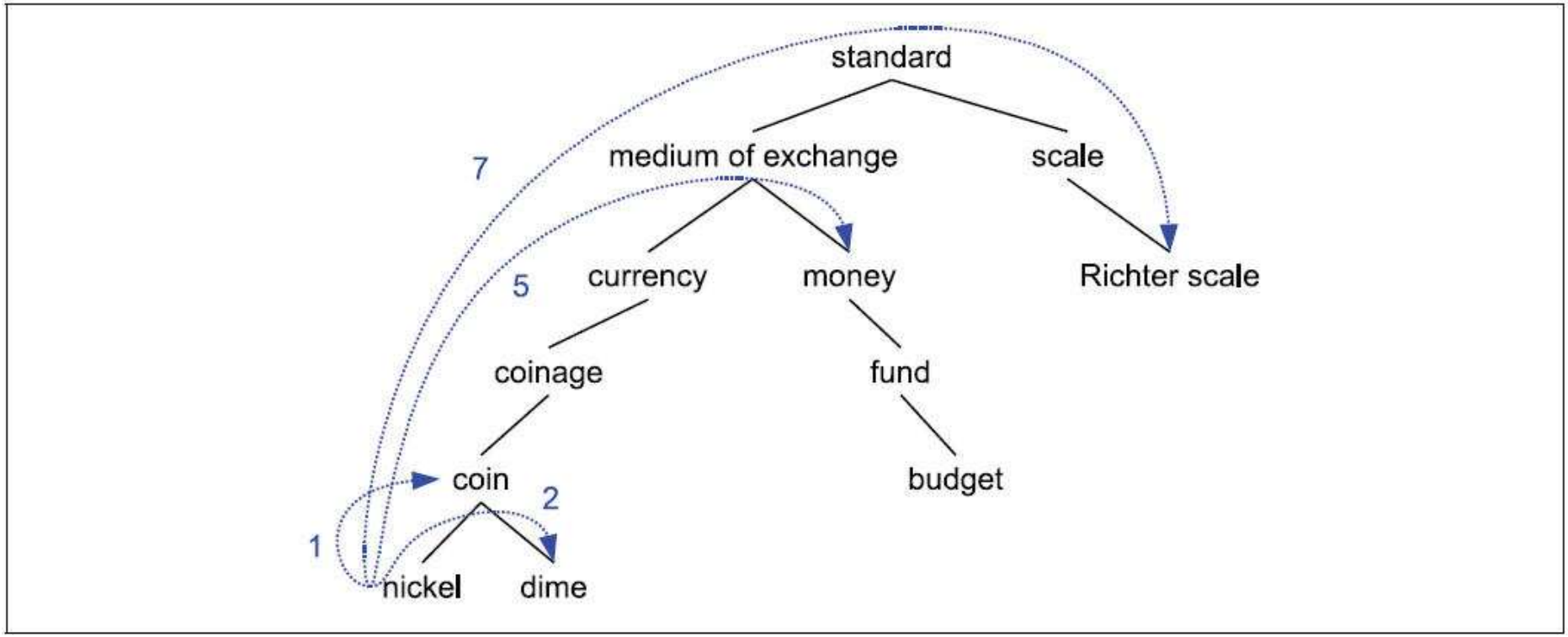
evergreen bearing foliage throughout the year

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

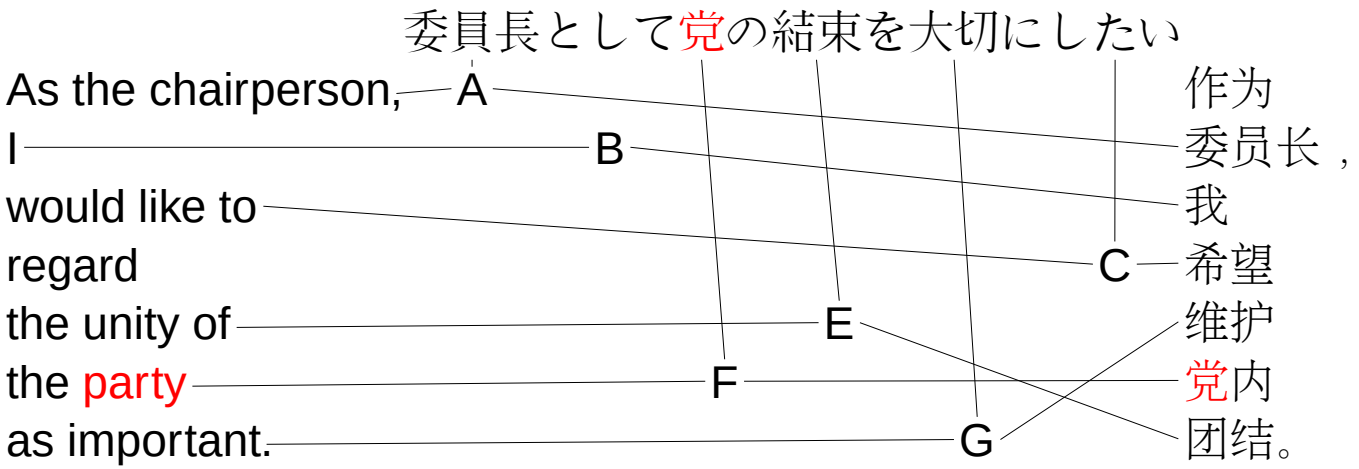
- First Sense: 63% (baseline)
- Extended Lesk: 68%
- Supervised: 70-72% (most words)
- Much harder task than POS tagging
  - Improve by reducing granularity (cluster senses)
  - Improve by increasing training data
  - Improve with more features (adding in syntax)

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.

Readings

- **Machine Translation**: Jurafsky and Martin (2009), Chapter 25.1–2
- **Word Sense Disambiguation**: Jurafsky and Martin (2009), Chapter 20.1–8
- Some slides based on Rada Mihalcea and Ted Pedersen's tutorial at AAAI-2005 "Advances in Word Sense Disambiguation"
- Nice demo of similarities at: `marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi`