

Dependency parsing and comprehensive collocation extraction

Pete Whitelock

Sharp Labs. of Europe

Oxford

Why?

- Sharp Corporation work on MT (1979 -)
 - Japanese to English
- SLE's Intelligent Dictionary (1997 -)
 - help Japanese speakers to read English (by glossing)
& support acquisition of English reading skills
- Obvious need for new application
 - Help Japanese to write English
 - Bilingual example retrieval
 - Context-sensitive thesaurus
 - Error detection AND correction

Types of Error

- Omissions, Insertions (usu. of grammatical elements)
 - Word Order
 - Replacements
 - Context-sensitive (real-word) spelling errors
 - Homophone and near-homophone errors
 - to/too/two, their/there/they're
 - lose/loose, pain/pane
 - lend/rend
 - Typos
 - bane/babe
 - than/that, from/form
 - Morphological errors
 - inflectional, eg agreement errors
 - derivational
 - safety/safe/safely
 - interested/interesting
-
- The diagram illustrates the classification of errors. Two labels, 'category-preserving' and 'category-changing', are positioned on the right side. Arrows point from these labels to various error types: 'category-preserving' points to 'Homophone and near-homophone errors', 'Typos', and 'Morphological errors'; 'category-changing' points to 'Context-sensitive (real-word) spelling errors' and 'derivational' errors.

Context-sensitive thesaurus

- How do we describe the following in positive terms: *a career; an impression; an achievement; a performance?*
- What's the right verb for making: *a relationship; a diversion; a model; a policy?*
- What's the right verb for using words to convey: *an opinion; a theory; the truth; an objective?*
- How do we describe extreme cases of: *a crime; a disagreement; a mistake; a change?*
- What feelings typically accompany: *enthusiasm; prejudice; fear; confidence?*

Semantic and collocational errors

- I don't want to marry with him. → 0
- We walked in the farm → on
- Then I asked 0 their suggestions → for
- I used to play judo but now I play karate → do
- Please teach me your phone number → tell/give
- Could you teach me the way to the station → tell
- I became to like him → came/started
- The light became dim → grew
- When he became 16 → turned/reached
- My boyfriend presented me some flowers → gave
- I'm very bad at writing pictures → drawing
- My brother always wins me at tennis → beats
- My father often smacked my hip → bottom
- Tradition in our dairy life → daily

Approaches to Error Detection

- Symbolic
 - IBM's Epistle (1982), Critique (1989)
 - Heidorn, Jensen, Richardson et al.
 - => MS Word Grammar Checker (1992→)
- Machine Learning
- Statistical

Symbolic Approaches

- IBM's Epistle (1982), Critique (1989)
 - Heidorn, Jensen, Richardson et al.

⇒ MS Word Grammar Checker (1992→)
- Full rule-based parsing
- Error rules ($S \rightarrow NP[+sg] VP[+pl]$)
- Confusion sets
 - eg *alter/altar, abut/about*
 - when one member appears, parse with all
 - only effective when POS's disjoint

Statistical Intuition

Given a comprehensive model of probability of bits of language,
improbable stretches of text correspond to errors

Can we build a comprehensive model?

Does the intuition hold?

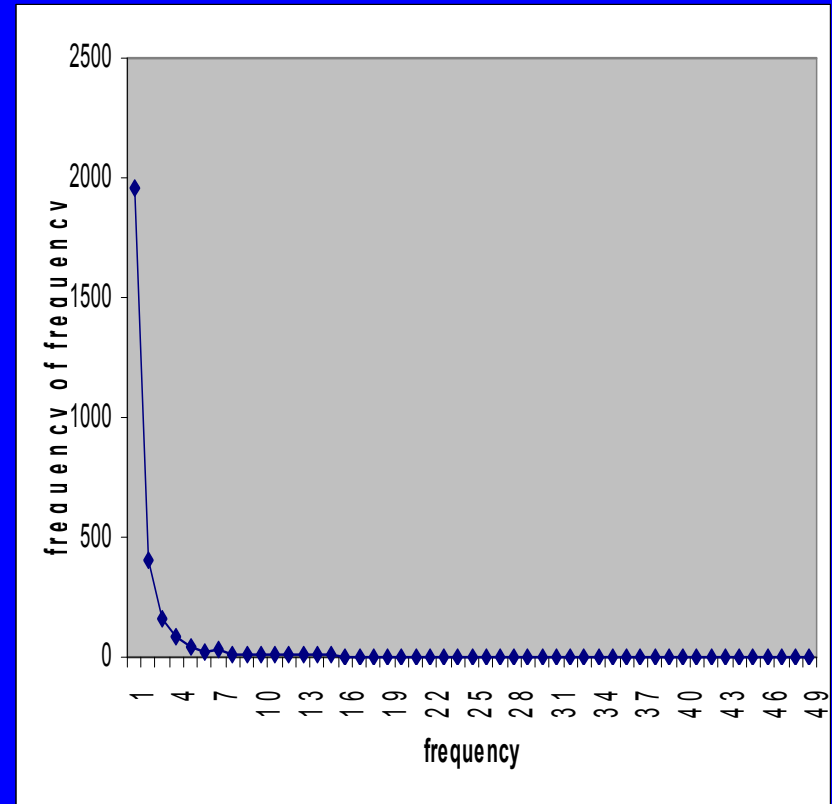
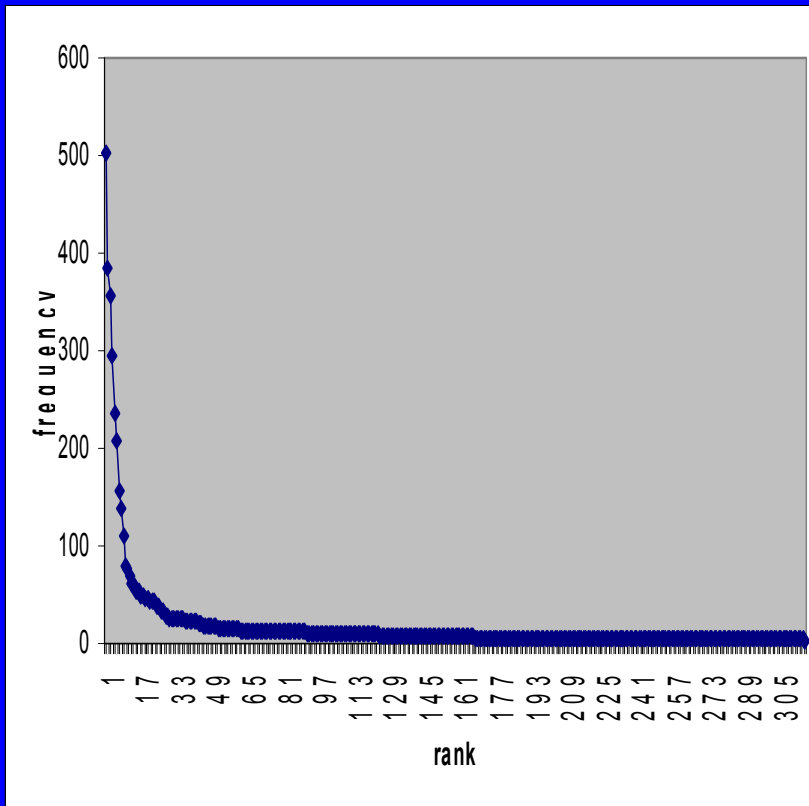
Statistical Approaches

Word n-grams

- IBM – Lange (1987), Damerau (1993)
 - based on success of ASR technology
 - severe data sparseness problems
 - Eg For a vocabulary of 20,000 words:

2	400 million
3	8 trillion
4	1.6×10^{17}

... worse still



Problem

- there are many tokens of rare types
- there are few types of common token

=> data sparseness

And ...

- Any given N is not enough:
 - ... must **cause** awfully bad **effect**
 - ... we **make** many specialised **magazines**

... but

- There are techniques to deal with data sparseness – smoothing, clustering, etc.
- Trigram model is very effective
- Especially when use POS n-grams

Statistical Approaches II

POS n-grams

- Atwell (1987), Schabes et al. (1996)

It is to fast.
PP BEZ PREP ADJ STOP

to_PREP confusable_with too_ADV

$p(\text{ADV ADJ STOP}) \gg$
 $p(\text{PREP ADJ STOP})$

Not appropriate for items with same POS

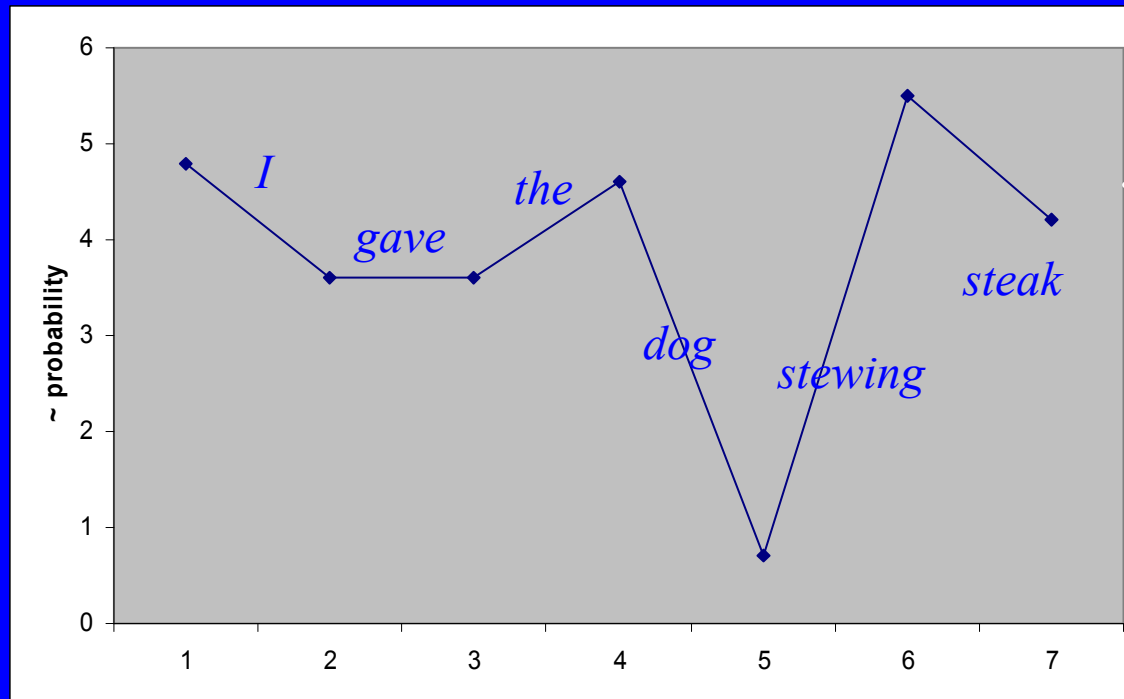
Machine Learning Approaches

- techniques from WSD
- define confusable sets C
- define features of context
 - eg specific words, POS sequences, etc
- learn map from features to elements of C
 - Bayes, Winnow (Golding, Schabes, Roth)
 - LSA (Jones & Martin)
 - Maximal entropy (Izumi et al.)
- effective, esp. for category preserving errors

Problems

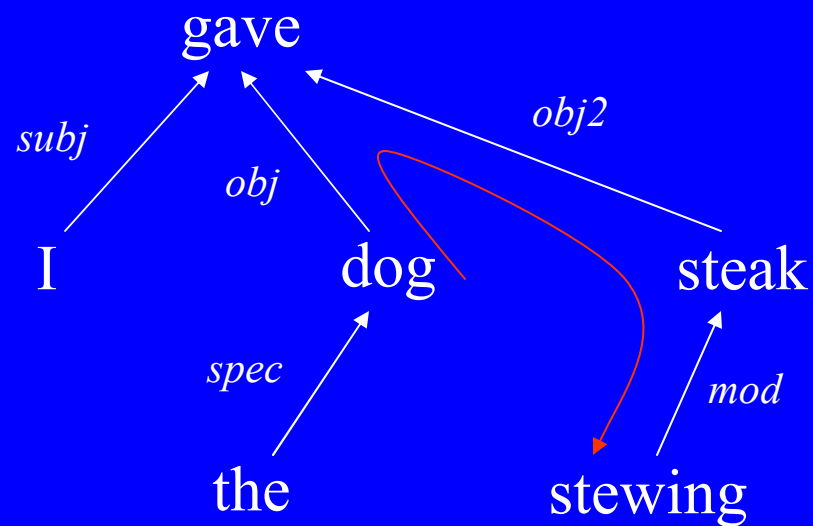
- experiments typically restricted to small set of spelling-type errors
- but almost any word can be used in error
- data problems with scaling up
- semantic-type errors have huge confusion sets
- but presence in a confusion set is the only trigger for error processing
- where is the probabilistic intuition?

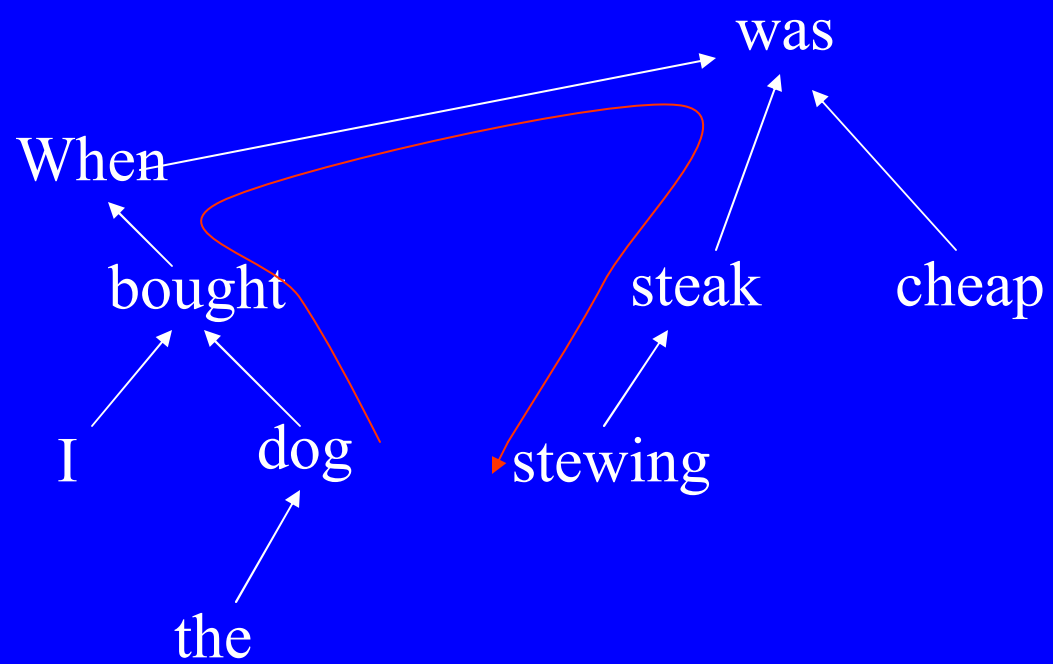
The Statistical Intuition - a problem



Dependency Structure

I gave the dog stewing steak





... SO

- Items that are linguistically close may be physically remote
 - ⇒ difficult to train n-gram model
- Items that are physically close may be linguistically remote
 - ⇒ low probabilities are sometimes uninteresting

ALEK

Chodorow & Leacock (2000)

- Compute MI for word bigrams and trigrams
- 30 words, 10,000 examples for each (NANC)
- TOEFL grade correlates significantly with proportion of low-frequency n-grams
- Mitigate uninteresting improbability by aggressive thresholding
- => v. low recall (c. 20%) for 78% precision

Bigert & Knutsson (2002)

- Detect improbable tri(pos)grams
- Use result of parsing to detect trigrams straddling syntactic boundaries, and ignore
=> mitigate uninteresting improbability

Idea

- Compute strength of links between words that are adjacent in dependency structure.
 - => Concentrate sparse data in linguistic equivalence classes
 - => Capture physically longer dependencies
 - => Weaker links should be a more reliable indicator of an error
 - => Error correction can be triggered only when required

Off-line

- parse a large quantity of text written by native speakers (80 million words BNC)
- produce dependency structures
- count frequencies of $\langle \text{word}_1, \text{dep}, \text{word}_2 \rangle$ types and compute strength
- build database of word combinations (collocations – free – anti-collocations)

Parsing I

- Get small quantity of hand tagged labeled bracketed text (1 million words Brown corpus)
- Exploit labeling to enrich tagset

```
( (S (NP Implementation/NN_H (PP of/OF (NP (NP Georgia/NP) 's/POS (NBAR automobile/NN_M title/NN_M) law/NN_H))) (AUX was/BED) (VP also/RB recommended/VBB_I (PP by/BY (NP the/AT outgoing/AJJ jury/NN_H)))) ./.)
```

Tagset

- AJJ attributive adjective
- PJJ predicative adjective
- NN_H head noun
- NN_M modifier noun
- AVBB attributive past participle
- AVBG attributive present participle

Tagset (cont.)

- VB(B|D|G|H|I|P|Z)_(I|T|A)
 - Verb forms for different transitivity
- BE(D|G|H|I|P|Z) copula
- HV(D|G|I|P|Z) auxiliary *have*
- DO(D|P|Z) auxiliary *do*
- MD modals
- TO infinitival *to*

Tagset (cont.)

- AT *a, the*
- DT demonstrative determiners (*this, each*)
- DP various pronouns (*this, nothing, each*)
- PP\$ possessive determiners
- SPP subject pronouns
- OPP object pronouns
- EX existential *there*
- SC subordinating conjunction
- PREP prepositions except *by* and *of*
- BY *by*
- OF *of*

Tagset (cont.)

- RB regular adverb
- RC 2nd as in ‘*as X as Y*’
- RD predet adverb (*only, just*)
- RG post-np adverb (*ago, aside, away*)
- RI pre-SC/PREP adverb (*only, even, just*)
- RJ pre-adjective adverb (*as, so, very, too*)
- RQ pre-numeral adverb (*only, about*)
- RT temporal adverb (*now, then, today*)
- NT temporal NP (*last week, next May*)

- Define dependency grammar in terms of enriched tags

```

ATVERB obj ACC_NP
MAIN_VERB vcompt TO # hope, expect etc.
MAIN_VERB vcomp_i INF_VERB # have, help, let, see etc.
MAIN_VERB vcomp_g GER # stop, start, catch, keep etc.
MAIN_VERB vcomp_b VBB_T # have, get
VBA vcomp_j PJJ0 # seem, become, feel
etc.

```

where:

```

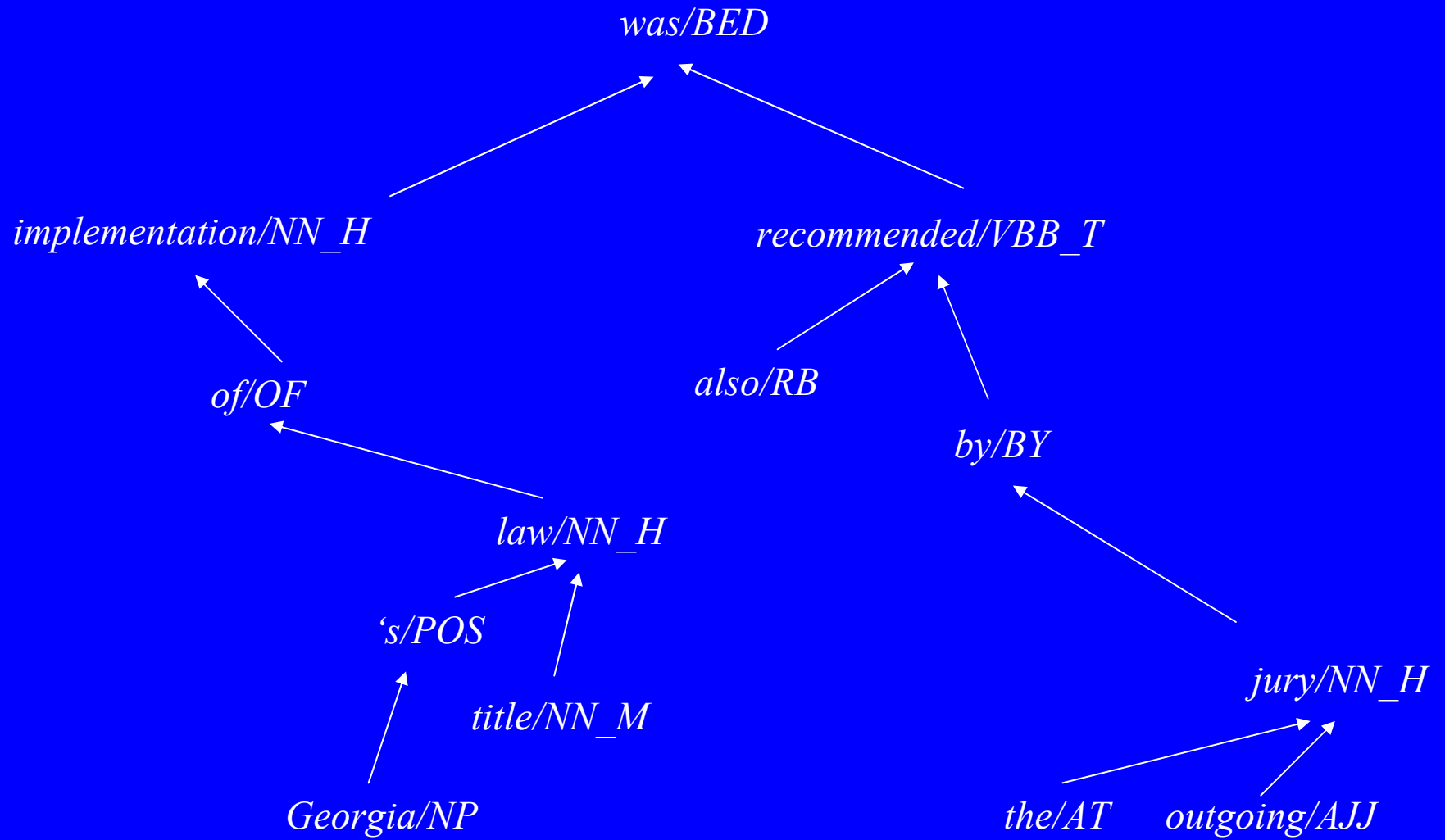
ATVERB = VBI_T|VBH_T|VBP_T|VBD_T|VBZ_T|VBG_T
VBA = VBD_A|VBI_A|VBH_A|VBP_A|VBZ_A
FIN_MAIN_VERB = VBZ_I|VBP_I|VBD_I|VBZ_T|VBP_T|VBD_T|VBZ_A|VBP_A|VBD_A
INF_MAIN_VERB = VBI_I|VBI_T|VBI_A
INF_VERB = BEI|HVI|INF_MAIN_VERB
P_PAR = BEH|VBH_I|VBH_T|VBH_A
GER = BEG|HVG|VBG_I|VBG_T|VBG_A
NFIN_VERB = INF_VERB|P_PAR|VBB_T|GER|TO
NFIN_MAIN_VERB =
    INF_MAIN_VERB|VBB_T|VBG_I|VBG_T|VBG_A|VBH_I|VBH_T|VBH_A
MAIN_VERB = FIN_MAIN_VERB|NFIN_MAIN_VERB
VERB = FIN_VERB|NFIN_VERB

```

define various sets of tags to use in statement of possible dependencies

- Exploit labeled bracketing to compute dependency structure

```
0:Implementation/NN_H<subj-7>  
1:of/OF<pmod-0>  
2:Georgia/NP<pos-3>  
3:'s/POS<spec-6>  
4:automobile/NN_M<mod-6>  
5:title/NN_M<mod-6>  
6:law/NN_H<pobj-1>  
7:was/BED  
8:also/RB<adv-9>  
9:recommended/VBB_T<ccompb-7>  
10:by/BY<padv-9>  
11:the/AT<spec-13>  
12:outgoing/AJJ<mod-13>  
13:jury/NN_H<pobj-10>  
14:./.
```



- compute MLE that two words with tags t_i and t_j , separated by n words, are in a dependency relation

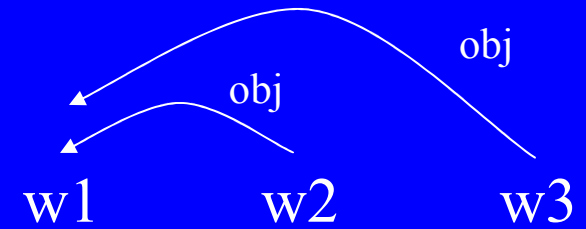
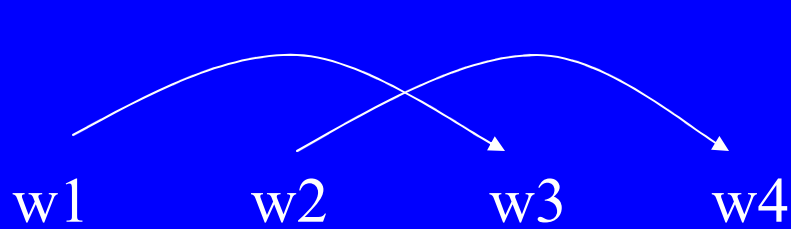
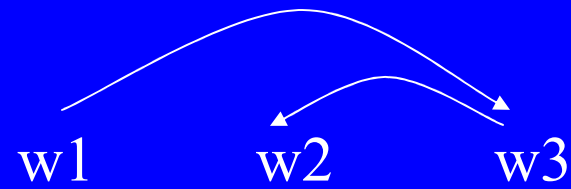
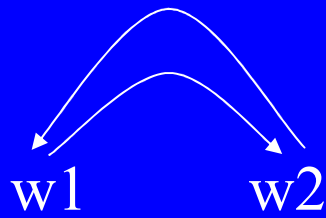
ltag	rtag	ltag_is	sep	rel	poss	actual	%
AT	NNS_H	D	1	spec	7207	6966	96
AT	NNS_H	D	2	spec	4370	4225	96
AT	NNS_H	D	3	spec	3204	1202	37
AT	NNS_H	D	4	spec	4300	325	7
AT	NNS_H	D	5	spec	4061	78	1

~36,000 entries

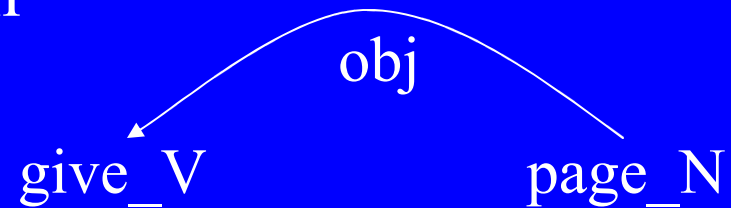
On the large corpus:

- tag the input text (assign a probability to each possible part-of-speech a word might have)
- look at each pair of words, and at each tag that they might have, and compute the probability that they are in a dependency relation with those tags at that distance apart
- sort the potential relations by probability
- apply greedy algorithm that tries to add each dependency in turn and checks that certain constraints are not violated
- stop adding links when threshold exceeded – initially high

Constraints



- count raw frequencies for each pair of lemmas in combination



compute contingency table:

	obj-Y	obj-page
X-obj	8002918	2103
give-obj	150854	10

Compute metric which normalises for frequency of elements (eg t-score)

$$\frac{f(\textit{give-obj-page}) / N - (f(\textit{give-obj}) \times f(\textit{obj-page})) / N^2}{\sqrt{f(\textit{give-obj-page}) / N^2}}$$

- if combination is more likely than chance, metric is positive
- if combination is less likely than chance, metric is negative

For example

$$t_{give-obj-page} = -9.4$$

$$t_{devote-obj-page} = +6.0$$

Metrics

- MI – overestimates infrequent links
- T – seems best for error spotting
- Yule's Q – easy to normalise, 'squashed'
- χ^2 (chi-squared) – unsigned, spread
- λ (log-likelihood) – spread

65 m link tokens

6.5 m types (+ 1.5 m trigrams) > 1

Bootstrapping

- initial high threshold means we don't find links like head of A in A N N and head of P in V N P N
- parse the large corpus again, adding a term to the probability calculation which represents the collocational strength (Q)
- set the threshold lower
- recompute collocational strength
- current parser (unlabeled) accuracy
~ 82% precision 88% recall

On-line

- (tag with learner tagger to deal with category-changing errors)
- parse 'learner' text according to the same grammar and compute strengths of all links
- sort links by weakness
- try replacing words in weakest link by confusables
- if link is strengthened, and other links are not significantly weakened
 - ⇒ suggest replacement
- repeat while there are links weaker than threshold

for instance:

associate with

beat me

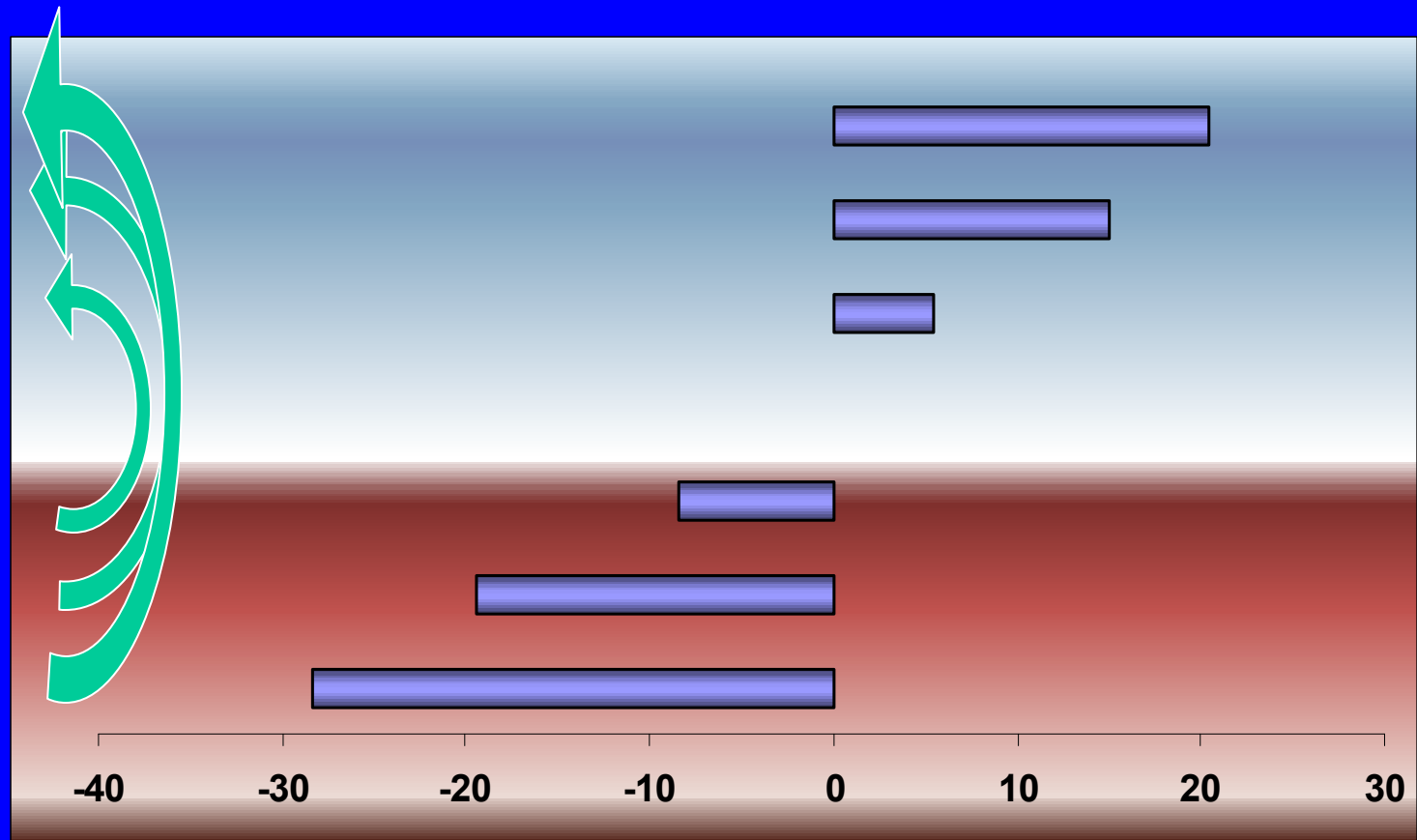
tall building

his property

high building

win me

associate to



t-score

Extend to 3+-grams

- by accident – GOOD
- by car accident – BAD

- a knowledge – BAD
- a knowledge of – GOOD

Data

- Development data: 121 Common Errors of English made by Japanese
- Training data: Brown/BNC (only for parser)
- Test data: extract from UCLES/CUP learner corpus (~3m words of exam scripts marked up with errors)

Confusables

- Annotations from learner corpus
- Co-translations from bilingual dictionary
- Synonym sets from thesaurus
 - published or derived from collocation data (Lin)
- Phonological neighbours
 - real-world spellos
 - use generalised edit distance with L1-L2 edits

Good Results

- We **settled** down **to** our new **house** ⇒
settled ... in ... house
- The **gases from cars** are ruining our atmosphere ⇒
emissions from cars
- Such experiments **caused** a bad **effect** ⇒
had ... effect
- We **had** a **promise** that we would visit the country ⇒
made ... promise
- I couldn't **study from** the evening **lecture** ⇒
learn from ... lecture
- It gives us the **utmost pleasure** ⇒ **greatest pleasure**

Bad Results:I

people say unlikely things

- Do you **remember** the **view** of sunrise in the desert? \Rightarrow
know view
- I listened to **every speech** \Rightarrow
every word
- Dudley's trousers slid down **his fat bottom** \Rightarrow
the bottom

SOLUTION: more data, longer n-grams

Bad Results:II

the input text is just too ill-formed

- I saw them who have **got** horrible injured **cause** of car accident ⇒
be cause

SOLUTION: Learner tagger

Bad Results:III

parser goes wrong

- My **most** disappointing **experience** ⇒
 great experience
- **Next**, the polluted **air** from the car does
 people harm ⇒
 close air

SOLUTION: Improve parser

Bad Results:IV

missed errors due to lack of evidence

- I will marry with my boyfriend next year
 - ‘marry with’ must be followed by one of small set of items – child, son, daughter
- I recommend you go interesting places
 - you can ‘go places’, but ‘places’ can’t be modified

SOLUTION: more data

Evaluation

	Preposition	Verb	Noun	Adjective
Total	576	475	242	115
Returned	232	165	89	37
Bad	42 (18.1%)	54 (32.7%)	26 (29.2%)	7 (18.9%)
Good(0)	62 (26.7%)	45 (27.3%)	19 (21.4%)	16 (43.2%)
Good(1)	83 (35.8%)	41 (24.9%)	37 (41.6%)	11 (29.7%)
Good(2-6)	37 (15.9%)	24 (14.5%)	7 (7.9)	3 (8.1%)
Good(7-)	8 (3.4%)	0 (0%)	0 (0%)	0 (0%)
Total Good	190	110	63	30
Precision	81.9%	66.7%	70.8%	81.1%
Recall	33.0%	23.2%	26.0%	26.1%

Summary of results

	PREP	VERB	NOUN	ADJ	Target	Cf. MS Word 2000
Precision	82%	67%	71%	81%	90%	~95%
Recall	33%	23%	26%	26%	25-50%	~5%

Conclusions and Directions

- finds and corrects types of error poorly treated in other approaches
- computing collocational strength is necessary but not sufficient for high precision, high recall error correction
- needs to be integrated with other techniques
- learn optimal combination of evidence eg by using collocational strengths as (some of) the features in a ML/WSD system
- deploy existing technology in other ways

An Explorable Model of English Collocation for
Writers, Learners, Teachers and Testers

<http://www.sle.sharp.co.uk/JustTheWord>