معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

The Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics

Preslav Nakov, Qatar Computing Research Institute
(joint work with Marti Hearst, UC Berkeley)

MWE'2014
April 26, 2014
Gothenburg, Sweden

Web-scale Computational Linguistics

The Big Dream
(*2001: A Space Odyssey*)

This is too hard!

So, we tackle sub-problems instead.

The Rise of Corpora

- The field was stuck for quite some time.
  - e.g., CYC: manually annotate all semantic concepts and relations

- A new statistical approach started in the 90s
  - Get **large** text collections.
  - Compute statistics over the words.

Size Matters

Banko & Brill: "Scaling to Very, Very Large Corpora for Natural Language Disambiguation", ACL'2001

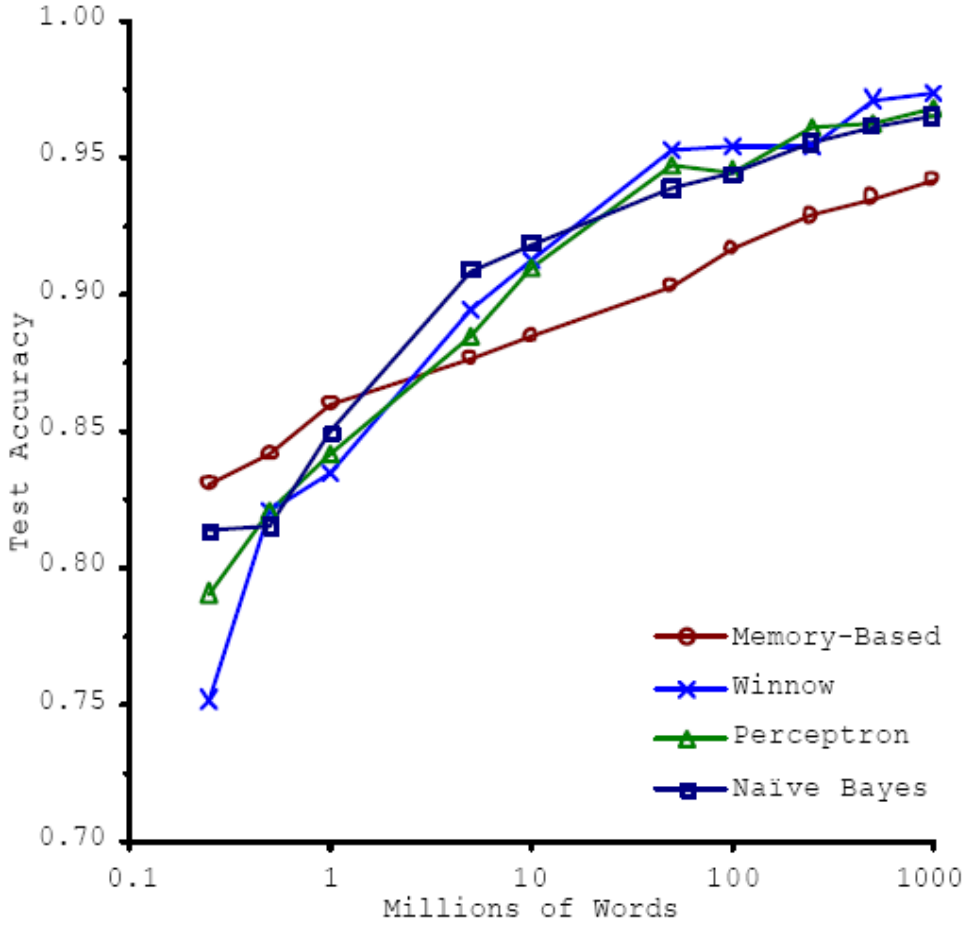
- **Spelling correction**
  - *Which word should we use?*

    <principal> <principle>

  - *In a given context:*
    - **Randy Evans is the Principal of Gothenburg School District 20.**
    - **Sweden's Foreign Minister declares his support for principles to protect privacy in the face of surveillance.**

Size Matters: Using *Billions* of Words

For this problem, one can get **a lot** of training data.



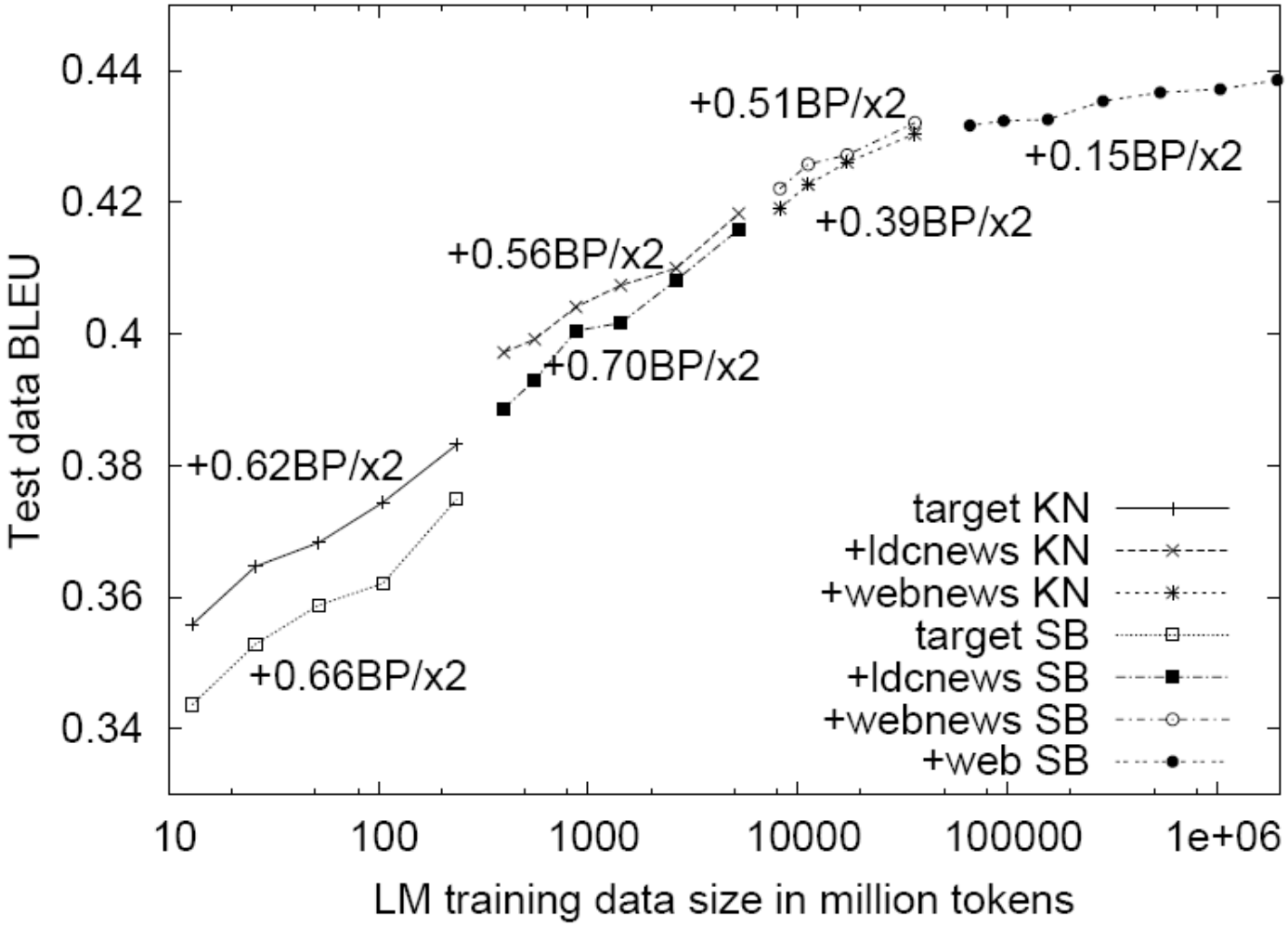
(Banko & Brill, 2001)

Great idea! Can it be extended to other tasks?

- Log-linear improvement even to a billion words!
- Getting more data is better than fine-tuning algorithms!

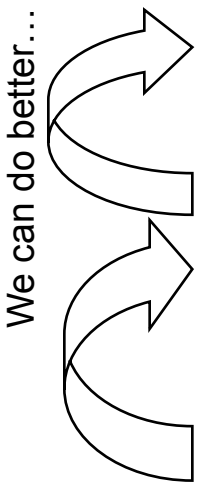
Language Models for SMT at Google: Using *Quadrillions (10^{15})* of Words!

(Brants&al,2007)



The Web as a Baseline

- "Web as a baseline" (Lapata & Keller 04;05): *n*-gram models



  - machine translation candidate selection
  - article generation
  - noun compound interpretation
  - noun compound bracketing
  - adjective ordering
  - spelling correction
  - countability detection
  - prepositional phrase attachment

Significantly better than the best supervised algorithm.

Not significantly different from the best supervised algorithm.

These are all UNSUPERVISED!

- Their conclusion:
  - **The Web should be used as a baseline.**

The Web as an Implicit Training Set

- Much more can be achieved using
  - surface features
  - paraphrases
  - linguistic knowledge

- I will demonstrate this on noun compounds

  (and on some other problems)

Noun Compounds

Noun Compound

- *Def: Sequence* of nouns that function as a *single* noun, e.g.
  - *healthcare reform*
  - *plastic water bottle*
  - *colon cancer tumor suppressor protein*
  - *Korpuslinguistikkonferenz (German)*

Three problems:
1. Segmentation
2. Syntax
3. Semantics

Noun Compounds

- Encode Implicit Relations – *hard to interpret*
  - *malaria mosquito* – CAUSE
  - *plastic bottle* - MATERIAL
  - *water bottle* - CONTAINER
- Abundant – *cannot be ignored*
  - 4% of the tokens in the Reuters corpus
- Highly productive – *cannot be listed in a dictionary*
  - 60.3% of the compounds in the British National Corpus occur just once
  - only 27% of English compounds of freq. >=10 are in an English-Japanese dictionary
- Also
  - ambiguous
  - context-dependent
  - (partially) lexicalized

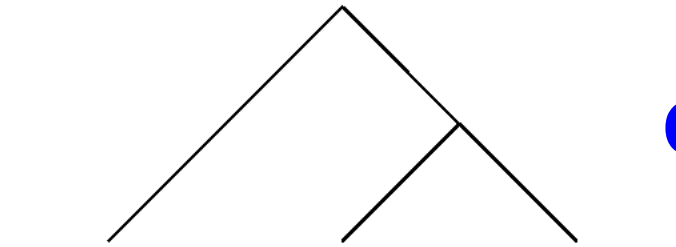
Noun Compounds: Applications

- **Question Answering, Machine Translation, Information Extraction, Information Retrieval**
  - *WTO Geneva headquarters* can be paraphrased as
    headquarters of the WTO located in Geneva
    Geneva headquarters of the WTO

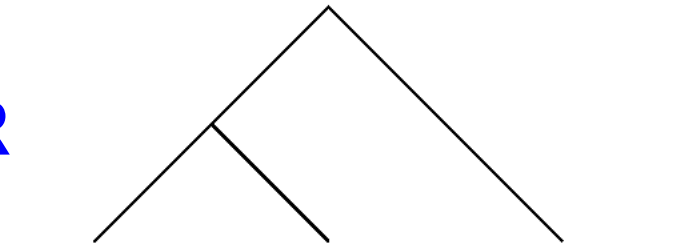
- **Information Retrieval**
  - Query: *migraine treatment*
  - verbs like *relieve* and *prevent* – for ranking and query refinement

Noun Compound Syntax

Noun Compound Syntax: The Problem



OR



?

[*plastic* [*water bottle*]]

right

water bottle made of *plastic*

[[*plastic water*] *bottle*]

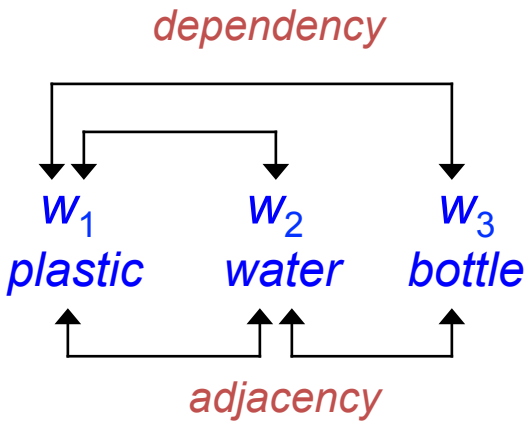
left

bottle containing *plastic water*

Measuring Word Association

Simple Word-based Models

- **Frequencies**
  - Dependency: #(w_1,w_2) vs. #(w_1,w_3)
  - Adjacency: #(w_1,w_2) vs. #(w_2,w_3)



- **Probabilities**
  - Dependency: Pr($w_1 \rightarrow w_2 | w_2$) vs. Pr($w_1 \rightarrow w_3 | w_3$)
  - Adjacency: Pr($w_1 \rightarrow w_2 | w_2$) vs. Pr($w_2 \rightarrow w_3 | w_3$)

- **Also: Pointwise Mutual Information, Chi Square, etc.**

Web-derived Surface Features

The Web as an Implicit Training Set

- Observations
  - Authors often disambiguate noun compounds using **surface markers**.
  - The size of the Web makes such markers frequent enough to be useful.
- Ideas
  - Look for instances where the compound occurs with ***surface markers***.
  - Also try
    - **paraphrases**
    - **linguistic knowledge**

Web-derived Surface Features: **Dash (hyphen)**

- Left dash
  - *cell-cycle analysis* ➔ *left*
- Right dash
  - *donor T-cell* ➔ *right*

Web-derived Surface Features: **Possessive Marker**

- After the first word
  - *world**'s** food production* ➔ *right*

- After the second word
  - *cell cycle**'s** analysis* ➔ *left*

Web-derived Surface Features: **Capitalization**

- don't-care – lowercase – uppercase
  - ***P****lasmodium* ***v****ivax* ***M****alaria* ➔ *left*
  - ***p****lasmodium* ***v****ivax* ***M****alaria* ➔ *left*

- lowercase – uppercase – don't-care
  - ***t****umor* ***N****ecrosis* ***F****actor* ➔ *right*
  - ***t****umor* ***N****ecrosis* ***f****actor* ➔ *right*

Web-derived Surface Features: **Embedded Slash**

- Left embedded slash
  - *leukemia/lymphoma cell* ➔ *right*

Web-derived Surface Features: **Parentheses**

- Single word
  - *growth factor (beta)* ➔ *left*
  - *(tumor) necrosis factor* ➔ *right*
- Two words
  - *(cell cycle) analysis* ➔ *left*
  - *adult (male rat)* ➔ *right*

Web-derived Surface Features: **Comma,dot,column,semi-column,…**

- Following the second word
  - *lung cancer**:** patients* ➔ *left*
  - *health care**,** provider* ➔ *left*

- Following the first word
  - home**.** health care ➔ *right*
  - *adult**,** male rat* ➔ *right*

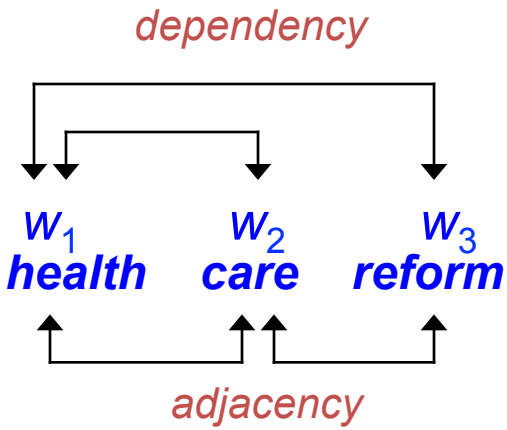
Web-derived Surface Features: **Abbreviation**

- After the second word
  - ***t****umor* ***n****ecrosis* ***(TN)*** *factor* ➔ *left*
- After the third word
  - *tumor* ***n****ecrosis* ***f****actor* ***(NF)*** ➔ *right*

Web-derived Surface Features: **Concatenation**

Consider "*health care reform*"

- Dependency model
  - *healthcare vs. healthreform*
- Adjacency model
  - *healthcare vs. carereform*



- Triples
  - *"healthcare reform" vs. "health carereform"*

Web-derived Surface Features: **Internal Inflection Variability**

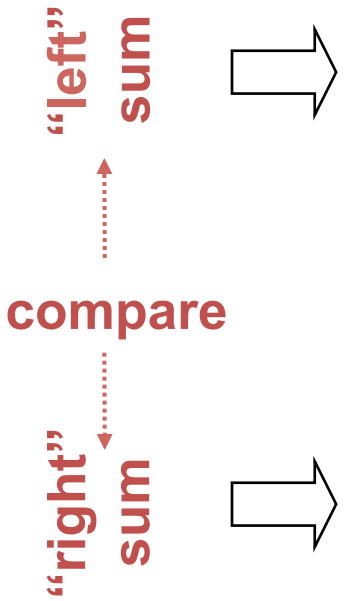
- First word
  - *bone mineral density*
  - *bone**s** mineral density*
- Second word
  - *bone mineral density*
  - *bone mineral**s** density*

Web-derived Surface Features: **Switch The First Two Words**

- Predict *right* if we can reorder
  - *adult male rat* as
  - *male adult rat*

Paraphrases

"bone marrow cell": left or right?



- **Prepositional**
  - *cells **in** (the) bone marrow* ➔ *left (61,700)*
  - *cells **from** (the) bone marrow* ➔ *left (16,500)*
  - *marrow cells **from** (the) bone* ➔ *right (12)*
- **Verbal**
  - *cells **extracted from** (the) bone marrow* ➔ *left (17)*
  - *marrow cells **found in** (the) bone* ➔ *right (1)*
- **Copula**
  - *cells **that are** bone marrow* ➔ *left (3)*

Evaluation Results

On 244 noun compounds from Grolier's encyclopedia (*Lauer dataset*)

- Word associations

Model	Correct	Wrong	N/A	Acc.	Cov.
χ^2 adjacency	184	60	0	75.41±5.77	100.00
χ^2 dependency	195	49	0	79.92±5.47	100.00

- Surface features and paraphrases

Model	Correct	Wrong	N/A	Acc.	Cov.
Concatenation adjacency	175	48	21	78.48±5.85	91.39
Concate					85.25
Concate					32.38
Inflection					43.03
Swap first					42.62
Reorder					62.30
Abbrevia					9.84
Possessives	32	4	208	88.89±14.20	14.75
Paraphrases	174	38	32	82.08±5.72	86.89
Surface features (sum)	183	31	30	85.51±5.34	87.70

Size does matter!

Using MEDLINE instead of the Web (million times smaller)

- 9.43% Coverage (23 out of 244 NCs)
- 47.83% Accuracy (12 out of 23 wrong)

CoNLL'05: Nakov&Hearst

Application to Other Syntactic Problems

Syntactic Application 1: Prepositional Phrase Attachment

(a) Peter spent millions of dollars. (*noun*)

(b) Peter spent time with his family. (*verb*)

Can be represented as a quadruple: (v, n1, p, n2)

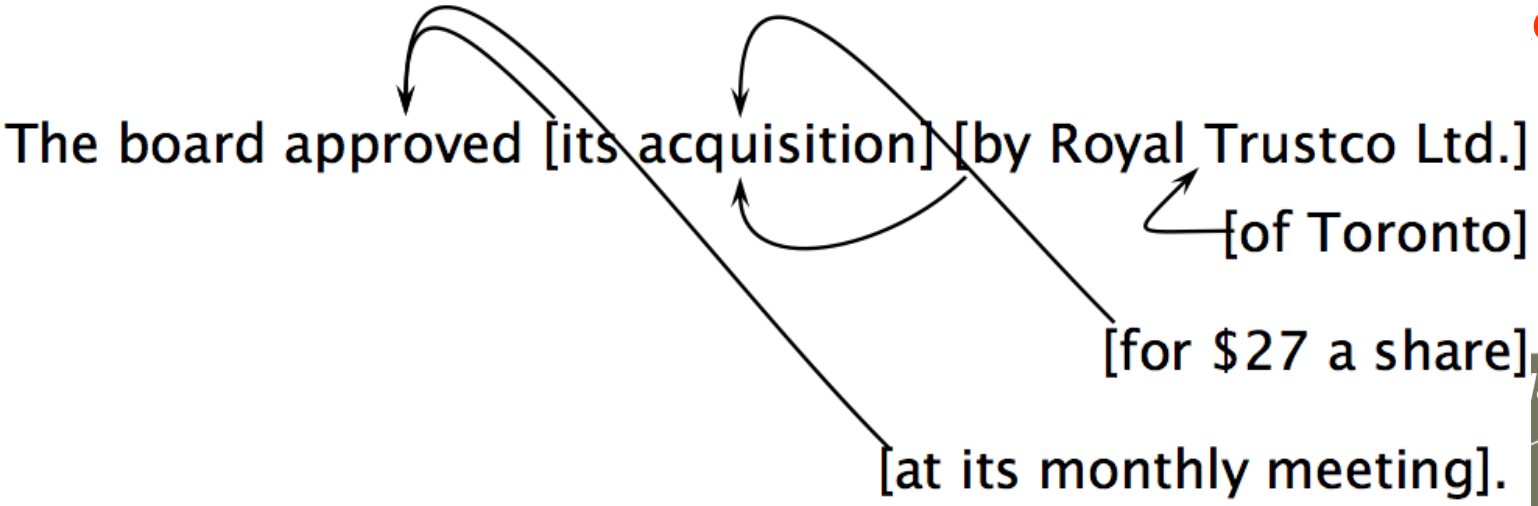
(a) (spent, millions, of, dollars)

(b) (spent, time, with, family)

Human performance:

- quadruple: 88%

ence: 93%



PP Attachment: *n*-gram models

- (*i*) Pr(p|n1) vs. Pr(p|v)
- (*ii*) Pr(p,n2|n1) vs. Pr(p,n2|v)
  - *I eat/v spaghetti/n1 with/p a **fork**/n2.*
  - *I eat/v spaghetti/n1 with/p **sauce**/n2.*

PP Attachment: Web-derived Surface Features

- Example features

	Acc	Cov
– *open the door / with a key* → verb	(100.00%	0.13%)
– *open the door (with a key)* → verb	(73.58%	2.44%)
– *open the door – with a key* → verb	(68.18%	2.03%)
– *open the door , with a key* → verb	(58.44%	7.09%)

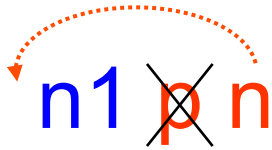
sum ⇦

compare

sum ⇦

	Acc	Cov
– *eat Spaghetti with sauce* → noun	(100.00%	0.14%)
– *eat ? spaghetti with sauce* → noun	(83.33%	0.55%)
– *eat , spaghetti with sauce* → noun	(65.77%	5.11%)
– *eat : spaghetti with sauce* → noun	(64.71%	1.57%)

PP Attachment Paraphrases (1)

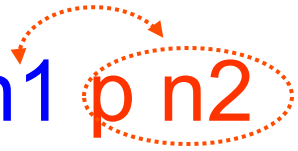


(1) v n1 ~~p~~ n2 ➔ v n2 n1 (noun)

- Turn "n1 p n2" into a noun compound "n2 n1"
  - *meet/v demands/n1 from/p customers/n2* ➔ *meet/v the customer/n2 demands/n1*

PP Attachment Paraphrases (2)

(2) v n1 p n2 ➔ v p n2 n1 (verb)



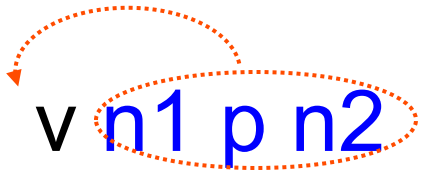
- Swap direct and indirect objects:
  - *had/v a program/n1 in/p place/n2* ➔
    had/v in/p place/n2 a program/n1

PP Attachment Paraphrases (3)

(3) v n1 p n2 ➔ p n2 * v n1 (verb)

- Look for apposition of "p n2"
  - *I gave/v an apple/n1 to/p him/n2* ➔
    (It was) to/p him/n2 (that) I gave/v an apple/n1

PP Attachment Paraphrases (4)



(4) v n1 p n2 ➔ n1 p n2 v (noun)

- Look for apposition of "n1 p n2"
  - *shaken/v confidence/n1 in/p markets/n2* ➔ *confidence/n1 in/p markets/n2 shaken/v*

PP Attachment Paraphrases (5)

(5) v n1 p n2 ➔ v PRONOUN p n2 (verb)

pronoun

- Substitute n1 with a pronoun (*him*, *her*)
  - *put/v a client/n1 at/p odds/n2* ➔
    *put/v **him** at/p odds/n2*

PP Attachment Paraphrases (6)

(6) v n1 p n2 ➔ BE n1 p n2 (noun)

to be

- Substitute v with *is/are/was/were*, e.g.
  - ***eat/v*** *spaghetti/n1 with/p sauce/n2* ➔
  - ***is*** *spaghetti/n1 with/p sauce/n2*

Syntactic Application 2: Noun Compound Coordination & Ellipsis

- Penn Treebank
  - ellipsis:

Real-world coordinations can be more complex:

*The Department of Chronic Diseases **and** Health Promotion leads **and** strengthens global efforts to prevent **and** control chronic diseases **or** disabilities **and** to promote health **and** quality of life.*

- Accuracy
  - Surface features & paraphrases: 80.61%

NP Coordination: N-gram models

(n1,c,n2,h)

- (*i*) #(n1,h) vs. #(n2,h)
- (*ii*) #(n1,h) vs. #(n1,c,n2)

NP Coordination: Surface Features

Example	Predicts	P(%)	R(%)
(buy) and sell orders	NO ellipsis	33.33	1.40
buy (and sell orders)	NO ellipsis	70.00	4.67
buy: and sell orders	NO ellipsis	0.00	0.00
buy; and sell orders	NO ellipsis	66.67	2.80
buy. and sell orders	NO ellipsis	68.57	8.18
buy[...] and sell orders	NO ellipsis	49.00	46.73
buy- and sell orders	ellipsis	77.27	5.14
buy and sell / orders	ellipsis	50.54	21.73
(buy and sell) orders	ellipsis	92.31	3.04
buy and sell (orders)	ellipsis	90.91	2.57
buy and sell, orders	ellipsis	92.86	13.08
buy and sell: orders	ellipsis	93.75	3.74
buy and sell; orders	ellipsis	100.00	1.87
buy and sell. orders	ellipsis	93.33	7.01
buy and sell[...] orders	ellipsis	85.19	18.93

NP Coordination Paraphrases (1)

(1) n1 c n2 h ➔ n2 c n1 h (ellipsis)

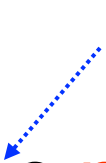
- Swap n1 and n2
  - bar/n1 and/c pie/n2 graph/h ➔ pie/n2 and/c bar/n1 graph/h

NP Coordination Paraphrases (2)

(2) n1 c n2 h ➔ n2 h c n1 (NO ellipsis)

- Swap n1 and n2 h
  - president/n1 and/c chief/n2 executive/h ➔ chief/n2 executive/h and/c president/n1

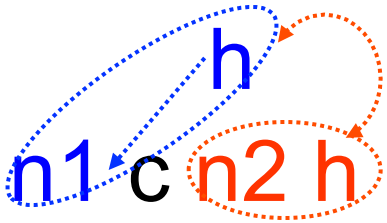
NP Coordination Paraphrases (3)



(3) n1 c n2 h ➔ n1 h c n2 h (ellipsis)

- Insert the elided head h
  - bar/n1 and/c pie/n2 graph/h ➔
  - bar/n1 graph/h and/c pie/n2 graph/h

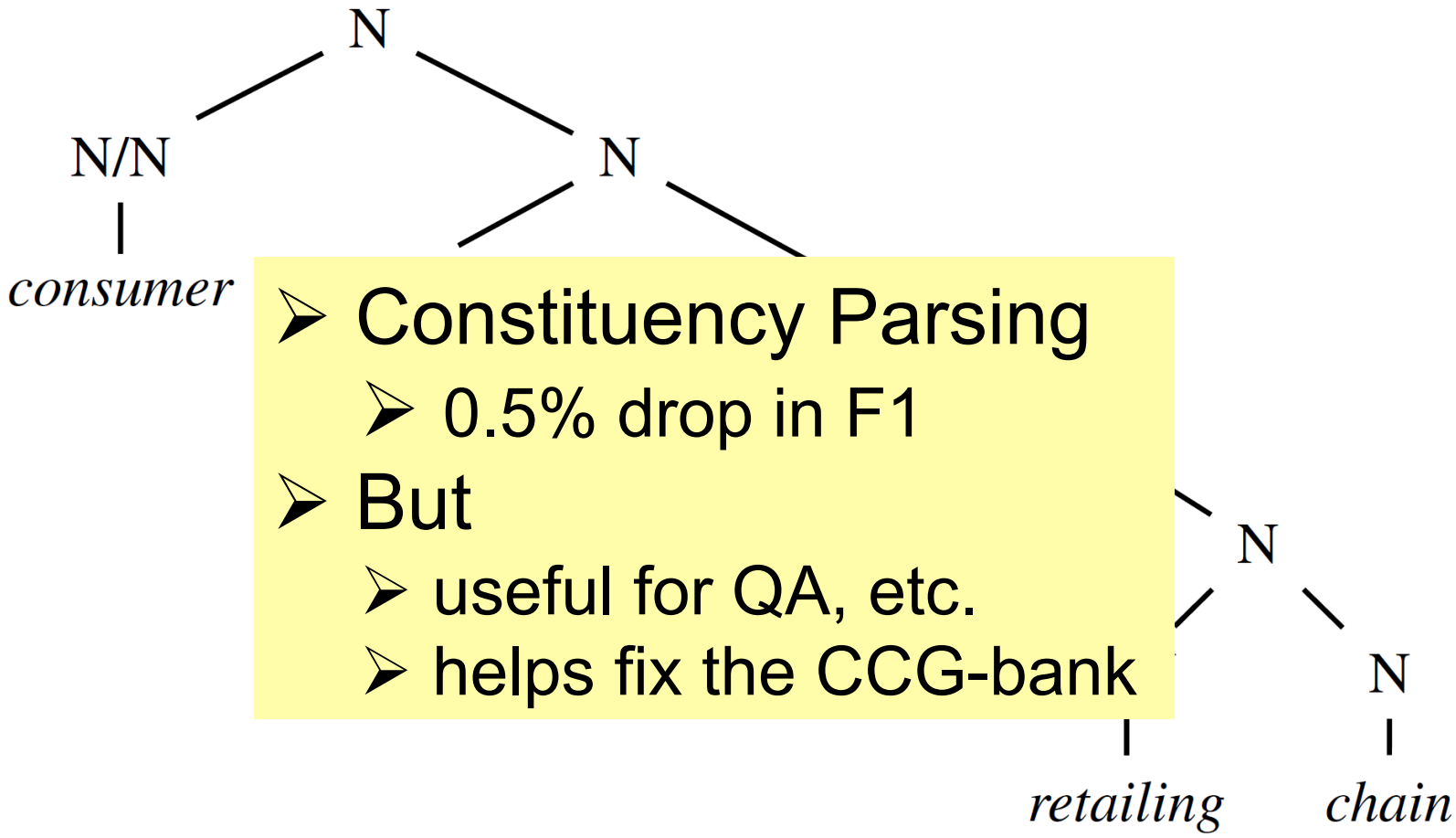
NP Coordination Paraphrases (4)



(4) n1 c n2 h ➔ n2 h c n1 h (ellipsis)

- Insert the head h; also switch n1 and n2
  - bar/n1 and/c pie/n2 graph/h ➔
  - pie/n2 graph/h and/c bar/n1 graph/h

More Applications (1): Bracketing the Penn Treebank NPs



- Constituency Parsing
  - 0.5% drop in F1
- But
  - useful for QA, etc.
  - helps fix the CCG-bank

Figure 1: CCG derivation from Hockenmaier (2003)

More Applications (2): Search Engine Query Segmentation

EMNLP'07: ***Learning Noun Phrase Query Segmentation***
Shane Bergsma and Qin Iris Wang

- Query Segmentation
  - [tumor suppressor protein]
  - [tumor suppressor] [protein]
  - [tumor] [suppressor protein]
  - [tumor] [suppressor] [protein]
- Bracketing

  [[tumor suppressor] protein]

  [tumor [suppressor protein]]

More Applications (3): Full Syntactic Parsing

ACL'11: *Web-Scale Features for Full-Scale Parsing*
Mohit Bansal and Dan Klein

POS_{head}	POS_{arg}	Example (*head, arg*)
RB	IN	*back → into*
NN	IN	*review → of*
NN	DT	
NNP	IN	
VB	NN	
VBD	NN	
NNP	NNP	
NN	TO	
JJ	IN	
NNS	TO	
IN	NN	
NNS	IN	*reports → on*
NN	NNP	*Warner ← studio*
NNS	JJ	*few ← plants*

POS_h	POS_a	mid/bfr-word	Example (*h, a*)
NNS	JJ	b = *the*	*other ← things*
NN	NN	m = -	*auto ← maker*
			adopted → plan
			omputer ← products
			the ← proposal
			going → into
			clusters → of
			In → review
			to → ease
			issue ← has
IN	NNS	m = *two*	*than → minutes*
IN	NN	b = *used*	*as → tool*
IN	VBD	m = *they*	*since → were*
VB	TO	b = *will*	*fail → to*

For constituency parsing, improvement is due to

- 40% affinity
- 60% paraphrases

Noun Compound Semantics

Noun Compound Semantics

- Typically, choose one abstract relation
  - **Fixed set of abstract relations** (Girju&al.,2005)
    - malaria mosquito ➔ CAUSE
    - olive oil ➔ SOURCE
  - **Prepositions** (Lauer,1995)
    - malaria mosquito ➔ with
    - olive oil ➔ from
- Proposed approach: use multiple paraphrasing verbs
  - **Paraphrasing verbs**
    - malaria mosquito ➔ carries, spreads, causes, transmits, brings, has
    - olive oil ➔ comes from, is obtained from, is extracted from
  - **Distribution over paraphrasing verbs**

Extracting Paraphrasing Verbs
Using a Linguistic Paraphrasing Pattern

pre-modifier

- Given "malaria mosquito", query Google for

"mosquito THAT * malaria"

post-modifier

- Extract verbs

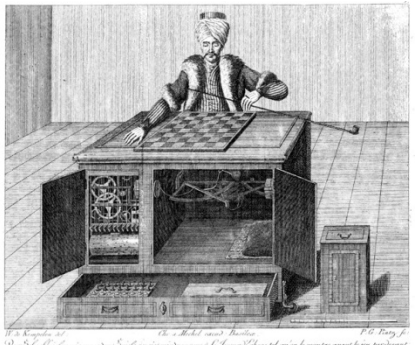
```
23 carry
16 spread
12 cause
 9 transmit
 7 bring
 4 have
 3 be infected with
 3 infect with
 2 give
```

Comparing to Girju&al. (2005)

Sem. relation	Example	Verbs extracted
POSSESSION	*"family estate"*	~~be in(29)~~, **be held by(9), be owned by(7)**
TEMPORAL	*"night flight"*	**arrive at(19), leave at(16), be at(6), be conducted at(6), occur at(5)**
IS-A (HYPERNYMY)	*"Dallas city"*	**include(9)**
CAUSE	*"malaria mosquito"*	*carry(23), spread(16),* **cause(12),** *transmit(9), bring(7), have(4),* *be infected with(3),* **be responsible for(3),** *test positive for(3),* **infect many with(3),** *be needed for(3),* **pass on(2), give(2), give out(2)**
MAKE/PRODUCE	*"shoe factory"*	**produce(28), make(13), manufacture(11)**
INSTRUMENT	*"pump drainage"*	**be controlled through(3), use(2)**
LOCATION/SPACE	*"Texas university"*	~~be(5)~~, **be in(4)**
PURPOSE	*"migraine drug"*	**treat(11), be used for(9), prevent(7), work for(6), stop(4), help(4),** ~~work(4)~~ **be prescribed for(3), relieve(3), block(3),** *be effective for(3), be for(3),* **help ward off(3),** *seem effective against(3),* **end(3), reduce(2), cure(2)**
SOURCE	*"olive oil"*	**come from(13), be obtained from(11), be extracted from(10),** **be made from(9), be produced from(7), be released from(4),** *taste like(4),* **be beaten from(3), be produced with(3), emerge from(3)**
TOPIC	*"art museum"*	**focus on(29),** *display(16),* **bring(14), highlight(11),** *house(10), exhibit(9)* **demonstrate(8), feature(7),** *show(5),* **tell about(4), cover(4), concentrate in(4)**
MEANS	*"bus service"*	**use(14), operate(6),** *include(6)*
EXPERIENCER	*"disease victim"*	**spread(12), acquire(12), suffer from(8), die of(7),** *develop(7),* **contract(6),** **catch(6), be diagnosed with(6),** *have(5), beat(5),* **be infected by(4), survive(4),** **die from(4), get(4), pass(3), fall by(3),** *transmit(3), avoid(3)*
THEME	*"car salesman"*	**sell(38),** ~~mean inside(13)~~, **buy(7), travel by(5), pay for(4), deliver(3), push(3)** **demonstrate(3), purr(3),** ~~bring used(3)~~, *know more about(3), pour through(3)*
RESULT	*"combustion gas"*	*support(22),* **result from(14), be produced during(11), be produced by(8),** **be formed from(8), form during(8), be created during(7), originate from(6),** **be generated by(6), develop with(6), come from(5),** ~~be cooled(5)~~

shown are 14 out of 21 relations

Amazon's Mechanical Turk: *Malaria Mosquito*



- 10 human judges:

➢ **The program:**

On 250 noun-noun compounds and 25-30 human judges: 32% cosine correlation

– **2 infects with**

– **1 has**

➢7 bring

SemEval-2010 task 9: ***The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions***

C. Butnariu, Su Nam Kim, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, T. Veale

– …

➢**2 give**

Relational Componential Analysis

- Classic componential analysis

	man	woman	boy	bull
ANIMATE	+	+	+	+
HUMAN	+	+	+	−
MALE	+	−	+	+
ADULT	+	+	−	+

- Relational componential analysis

	"cancer treatment"	"migraine treatment"	"wrinkle treatment"	"herb treatment"
treat	+	+	+	−
prevent	+	+	−	−
cure	+	−	−	−
reduce	−	+	+	−
smooth	−	−	+	−
contain	−	−	−	+
use	−	−	−	+

Noun Compound Semantics Using Abstract Relations

[colon cancer] [[tumor suppressor] protein]

ABSTRACT RELATIONS:

[
  [colon cancer]/LOCATION
  [[tumor suppressor]/PURPOSE protein]/AGENT
]/LOCATION

Noun Compound Semantics Using Prepositional Paraphrases

[colon cancer] [[tumor suppressor] protein]

PREPOSITIONS:

{

{protein that ***is*** a

{suppressor ***of*** tumors}

}

in

{cancer ***of/in*** the colon}

}

Noun Compound Semantics Using Paraphrasing Verbs

[colon cancer] [[tumor suppressor] protein]

VERBS:

{
{protein that ***acts as*** a
{suppressor that ***inhibits*** tumors}
}
which is ***implicated in***
{cancer that ***occurs in*** the colon}
}

prevent/stop/keep from developing/growing/arising

Free Paraphrasing of Noun Compounds: Going Beyond Verbs and Prepositions

"onion tears"

tears from onions

tears *due to cutting* onion

tears *induced when cutting* onions

tears *that* onions *induce*

tears *that come from chopping* onions

tears *that sometimes flow when* onions *are chopped*

tears *that raw* onions *give you*

SemEval-2013 task 4: Free Paraphrases of Noun Compounds

C. Butnariu, I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, T. Veale

Application to Other Semantic Tasks

The V+P+C Semantic Vector

- For "*noun1 noun2*", query:

  `"noun2 * noun1"`

  `"noun1 * noun2"`

- Extract:
  - V: verbs
  - P: prepositions
  - C: coordinating conjunctions

committee member

Freq.	Feature	POS	Direction
2205	of	P	2 → 1
1923	be	V	1 → 2
771	include	V	1 → 2
382	serve on	V	2 → 1
189	chair	V	2 → 1
189	have	V	1 → 2
169	consist of	V	1 → 2
148	comprise	V	1 → 2
106	sit on	V	2 → 1
81	be chaired by	V	1 → 2
78	appoint	V	1 → 2
77	on	P	2 → 1
66	and	C	1 → 2
66	be elected	V	1 → 2
58	replace	V	1 → 2
48	lead	V	2 → 1
47	be intended for	V	1 → 2
45	join	V	2 → 1
...	...	...	...
4	be signed up for	V	2 → 1

Semantic Application 1:
Predicting Levi's 12 Recoverably Deletable Predicates

RDP	Example	Subj/obj	Traditional Name
$\texttt{CAUSE}_1$	*tear gas*	object	causative
$\texttt{CAUSE}_2$	*drug deaths*	subject	causative
$\texttt{HAVE}_1$	*apple cake*	object	possessive/dative
$\texttt{HAVE}_2$	*lemon peel*	subject	possessive/dative
$\texttt{MAKE}_1$	*silkworm*	object	productive/composit.
$\texttt{MAKE}_2$	*snowball*	subject	productive/composit.
USE	*steam iron*	object	instrumental
BE	*soldier ant*	object	essive/appositional
IN	*field mouse*	object	locative
FOR	*horse doctor*	object	purposive/benefactive
FROM	*olive oil*	object	source/ablative
ABOUT	*price war*	object	topic

- Accuracy (212 noun-noun compounds)
  - V+P+C: 50.0%±6.7% (baseline: 19.6%)

Semantic Application 2:
SAT Analogy Questions

	ostrich:bird		**palatable:toothsome**
(a)	*lion:cat*	(a)	rancid:fragrant
(b)	goose:flock	*(b)*	chewy:textured
(c)	ewe:sheep	(c)	*coarse:rough*
(d)	cub:bear	(d)	solitude:company
(e)	primate:monkey	(e)	no choice

- Accuracy (174 noun:noun examples)
  - LRA : 67.4%±7.1%
  - V+P+C : 71.3%±7.0%

Semantic Application 3:

Relations Between Complex Nominals

SemEval-2007 task 4:Classification of semantic relations between nominals

R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, D. Yuret

"Among the contents of the <e1>vessel</e1> were a set of carpenter's <e2>tools</e2>, several large storage jars, ceramic utensils, ropes and remnants of food, as well as a heavy load of ballast stones."

WordNet(e1) = "vessel%1:06:00::",
WordNet(e2) = "tool%1:06:00::",
Content-Container(e2, e1) = "true",
Query = "contents of the * were a"

#	Relation Name
1	Cause-Effect
2	Instrument-Agency
3	Product-Producer
4	Origin-Entity
5	Theme-Tool
6	Part-Whole
7	Content-Container

- Accuracy
  - Task #4 winner : 66.0%
  - V+P+C : 67.0%
  - + web-based argument generalization : 71.3%

***Follow-up:* SemEval-2010 task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals**

I. Hendrickx, Su Nam Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz

Semantic Application 4:

30 Head-Modifier Relations

PARTICIPANT			
agent	ag	student protest	M performs H, M is animate or natural phenomenon
beneficiary	ben	student discount	M benefits from H
instrument	inst	laser printer	H uses M
object	obj	metal separator	M is acted upon by H
object property	obj_prop	sunken ship	H underwent M
part	part	printer tray	H is part of M
possessor	posr	national debt	M has H
property	prop	blue book	H is M
product	prod	plum tree	H produces M
source	src	olive oil	M is the source of H
stative	st	sleeping dog	H is in a state of M
whole	whl	daisy chain	M is part of H
QUALITY			
container	cntr	film music	M contains H
content			tained in H
equative			M
material			de of M
measure			easure of H
topic			cerned with M
type	type	oak tree	M is a type of H

- **Accuracy (600 examples)**
  - **LRA : 39.8%±3.8%**
  - **V+P+C : 40.5%±3.9%**

Semantic Application 5:
Textual Entailment

```
<pair id="284" entailment="YES" task="QA">
<t>While preliminary work goes on at the Geneva headquarters of the WTO,
  with members providing input, key decisions are taken at the ministerial
  meetings.</t>
<h>The WTO headquarters are located in Geneva.</h>
</pair>
```

"WTO Geneva headquarters" =

"headquarters of the WTO are located in Geneva"

(1) *Geneva headquarters of the WTO*

(2) *WTO headquarters are located in Geneva*

Application to Machine Translation

Statistical Machine Translation: Trained on Parallel Text

1 Europe's Divided Racial House
2 A common feature of Europe's e
of the immigration issue as a
3 The Lega Nord in Italy, the Vl
supporters of Le Pen's Nationa
of parties or movements formed
immigrants and promotion of si
4 While individuals like Jorg Ha
and (never to soon) go, the ra
European politics anytime soon
5 An aging population at home ar
increasing racial fragmentatic
6 Mainstream parties of the cent
confronted this prospect by hi
hoping against hope that the p
7 It will not, as America's raci
8 Race relations in the US have
the center of political debate
cleavages are as important as
determinants of political pref
9 The first step to address raci

1 La Dividida Cámara Racial de
2 Una característica común de
su racismo y su uso del tema
política.
3 La Lega Nord de Italia, el V
partidarios del Frente Nacio
ejemplos de partidos o movim
tema común de la aversión a
de políticas simplistas para
4 Aunque los individuos como J
vienen y (nunca demasiado pr
raza no desaparecerá de la p
momento cercano.
5 La población cada vez más vi
abiertas que nunca, implican
los países europeos.
6 Los principales partidos de
derecha se han enfrentado a
cabeza en la tierra, abrigan
problema desaparezca.
7 No lo hará, como claramente

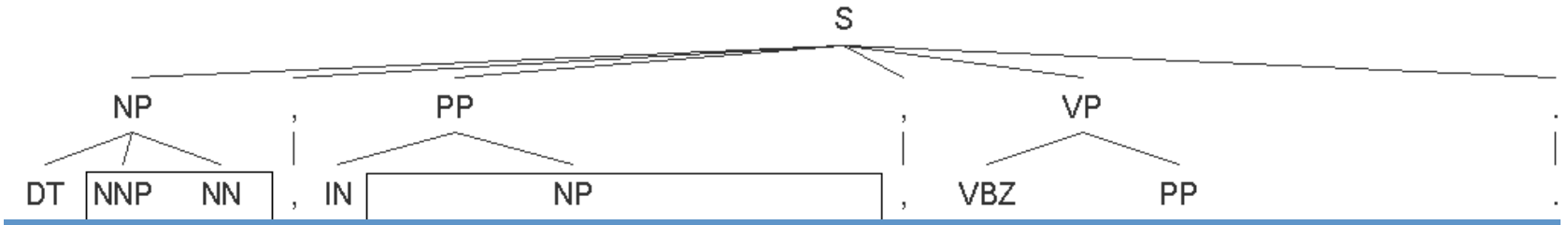
Noun Compounds in Phrase-based SMT

Idea: paraphrase the source phrase to increase coverage

oil price hikes ➔ alzas en los precios del petróleo
hikes *in* oil prices ➔ alzas en los precios del petróleo
hikes *in* prices *of* oil ➔ alzas en los precios del petróleo
hikes *in* prices *for* oil ➔ alzas en los precios del petróleo
hikes *in the* prices *of* oil ➔ alzas en los precios del petróleo
hikes *in the* prices *for* oil ➔ alzas en los precios del petróleo

oil price hikes	*alzas en los precios del petróleo*
of 1974 and 1980 ,	de 1974 y 1980 ,
Japan's economy	*la economía nipona*
recovered through	se recuperó a través del
export growth	*crecimiento basado en las exportaciones*
.	.

Paraphrasing a Source-Language Sentence



Improvement: equivalent to 33-50% of what could be achieved by doubling the amount of training data.

Looking forward to at least two papers on noun compounds in MT at this MWE'14:

- **German Compounds and Statistical Machine Translation. Can they get along?**
  Carla Parra Escartín, Stephan Peitz and Hermann Ney
- **Paraphrasing Swedish Compound Nouns in Machine Translation**
  Edvin Ullman and Joakim Nivre

The *budget of the EU* , as an economic policy instrument , amounts to 1.25 % of European GDP .
The *budget of the EU* , as an economic policy 's instrument , amounts to 1.25 % of European GDP .

Paraphrasing the Phrase Table

1	**% of members of the irish parliament** % of irish parliament members % of irish parliament 's members
2	**universal service of quality .** universal quality service . quality universal service . quality 's universal service .
3	**action at community level** community level action
4	**, and the aptitude for communication and** , and the communication aptitude and
5	**to the fall-out from chernobyl .** to the chernobyl fall-out .
6	**flexibility in development - and quick** development flexibility - and quick
7	**, however , the committee on transport** , however , the transport committee
8	**and the danger of infection with aids** and the danger of aids infection and the aids infection danger and the aids infection 's danger

Paraphrasing NPs & Noun Compounds

purely syntactic

1. $[_{\mathrm{NP}}\ \mathrm{NP}_1\ \mathrm{P}\ \mathrm{NP}_2] \Rightarrow [_{\mathrm{NP}}\ \mathrm{NP}_2\ \mathrm{NP}_1]$.
   the lifting of the beef import ban ⇒
   the beef import ban lifting.
2. $[_{\mathrm{NP}}\ \mathrm{NP}_1\ \textit{of}\ \mathrm{NP}_2] \Rightarrow [_{\mathrm{NP}}\ \mathrm{NP}_2\ \mathrm{poss}\ \mathrm{NP}_1]$.
   the lifting of the beef import ban ⇒
   the beef import ban's lifting.
3. $\mathrm{NP}_{poss} \Rightarrow \mathrm{NP}$.
   Commissioner's statement ⇒
   Commissioner statement.
4. $\mathrm{NP}_{poss} \Rightarrow \mathrm{NP}_{PP_{of}}$.
   Commissioner's statement ⇒
   statement of (the) Commissioner.

use Web statistics

5. $\mathrm{NP}_{NC} \Rightarrow \mathrm{NP}_{poss}$.
   inquiry committee chairman ⇒
   inquiry committee's chairman.
6. $\mathrm{NP}_{NC} \Rightarrow \mathrm{NP}_{PP}$.
   the beef import ban ⇒
   the ban on beef import.

Paraphrasing Noun Compounds

- Split the noun compound
  - N1="beef", N2="import ban lifting"
  - N1="beef import", N2="ban lifting"
  - N1="beef import ban", N2="lifting"

"lt N_1 poss N_2 rt"
"lt N_2 prep det N_1' rt"
"lt N_2 that be det N_1' rt"
"lt N_2 that be prep det N_1' rt"

- lt=word before
- rt=word after

Summary

Summary

- **Syntactic Tasks**
  - Noun Compound Syntax
  - Prepositional Phrase Attachment
  - Noun Compound Coordination
  - Full syntactic parsing, etc.
- **Semantic Tasks**
  - Noun Compound Semantics
  - Predicting
    - Abstract Semantic Relations
    - Relations Between Complex Nominals
    - Head-Modifier Relations
  - Solving SAT Analogy Problems
- **Application to a Real-World Task**
  - Machine Translation

Tapped the potential of very large corpora for corpus linguistics by going beyond the n-gram:

- surface markers
- paraphrases
- linguistic knowledge

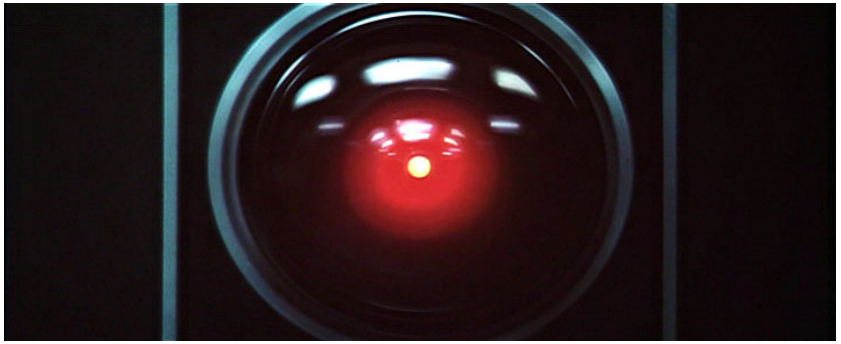
Some Useful Tools and Resources

- Yahoo! BOSS
- Google 1T 5-gram corpus
- Microsoft Web N-gram services
- IBM Web Fountain
- WaCKy
- Sketch engine

Future Directions

The Big Dream
(*2001: A Space Odyssey*)

Dave Bowman: “Open the pod bay doors, HAL”



HAL 9000: “I’m sorry Dave. I’m afraid I can’t do that.”

Semantics: Revolution is Needed?

- If we want the dream come true, we should
  - not rely on superficial statistics alone
  - need to get to the *meaning* of text

- A revolution in semantics is needed
  - looking at *words* is not enough
  - we need better models for
    - **multi-word expressions** (~70% of terminology)
    - **semantic relations (meaning is in the links!)**

- Key elements (in my opinion)
  - Web-scale corpora
  - linguistic knowledge
  - paraphrases

"Moving Lexical Semantics from Alchemy to Science"

Recent discussion on [Corpora-List]

- ***This is what Chomsky has done with syntax.***
- ***Should we expect the same for lexical semantics?***

Semantics: Community Efforts

- **Evaluations on shared corpora**
  - SemEval *(18 tasks in 2015: fragmentation or community expansion?)*
  - Shared tasks at *SEM and workshops
- **Special journal issues**
  - Computational Linguistics, LRE, JNLE, etc.
- **Workshops**
  - **Really, really fragmented!**
    - MWE, RELMS, Disco, GEMS, TextInfer,…
  - **But now we also have *SEM!**
  - **And established workshops such as MWE:**
    - 2-day, 10-years old, …
    - MWE section of SIGLEX

The Future?

Three words: Web, *paraphrases, linguistics*