The Web as a Corpus: Going Beyond the n-gram

Preslav Nakov

National University of Singapore

(joint work with Marti Hearst, UC Berkeley)

University of Heidelberg

February 3, 2011

Plan

- Introduction
- Surface Features & Paraphrases
- Syntactic Tasks
- Semantic Tasks
- Application to Machine Translation

Introduction

NLP: The Dream

This is too hard!

So, we tackle sub-problems instead.

How to Tackle the Problem?

- 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 '01: "Scaling to Very, Very Large Corpora for Natural Language Disambiguation", ACL

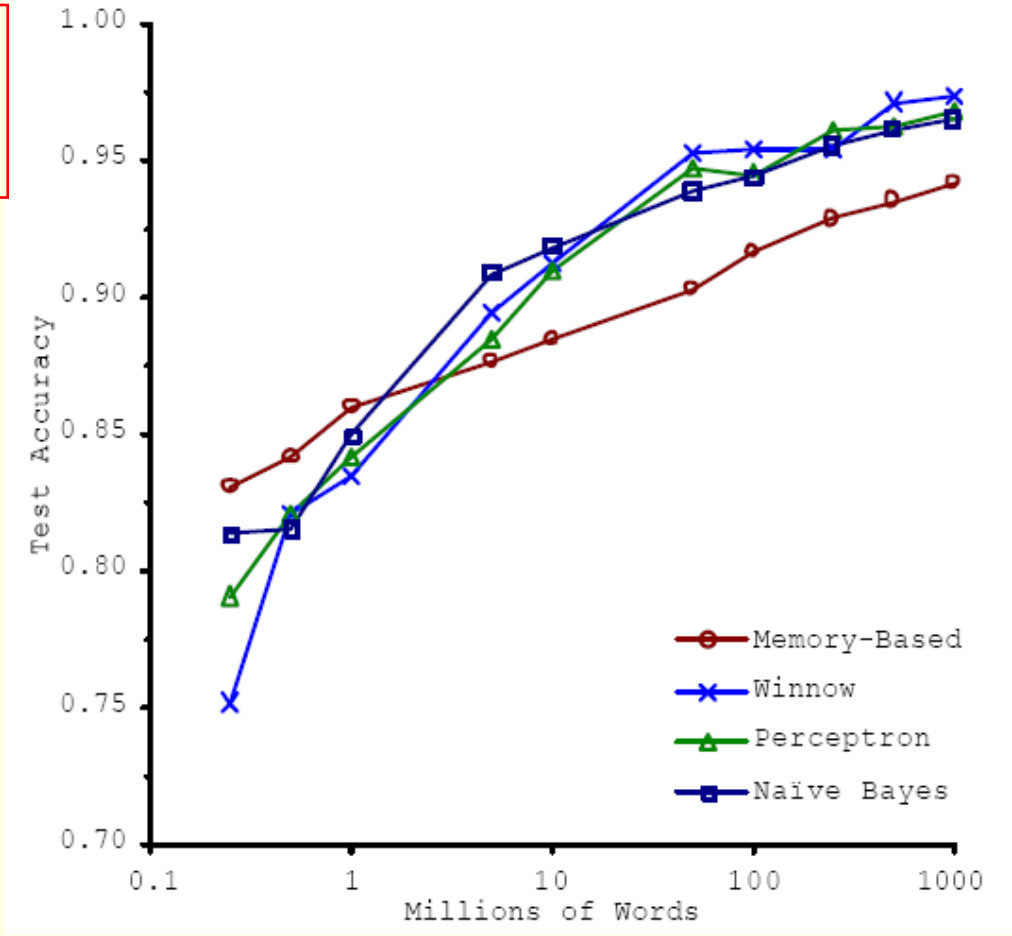
- Spelling correction: Which word should we use?

  <principal> <principle>

- Use context:
  - I am in my third year as the principal of Anamosa High School.
  - Power without principle is barren, but principle without power is futile. (Tony Blair)

Bigger is better than smarter!

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



Banko & Brill '01

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!

Web as a Baseline

- "Web as a baseline" (Lapata & Keller 04;05): applied simple *n*-gram models to
  - 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 than the best supervised algorithm.

These are all UNSUPERVISED!

- **Their conclusion:**
  - **=> Web *n*-grams should be used as a baseline.**

Contribution

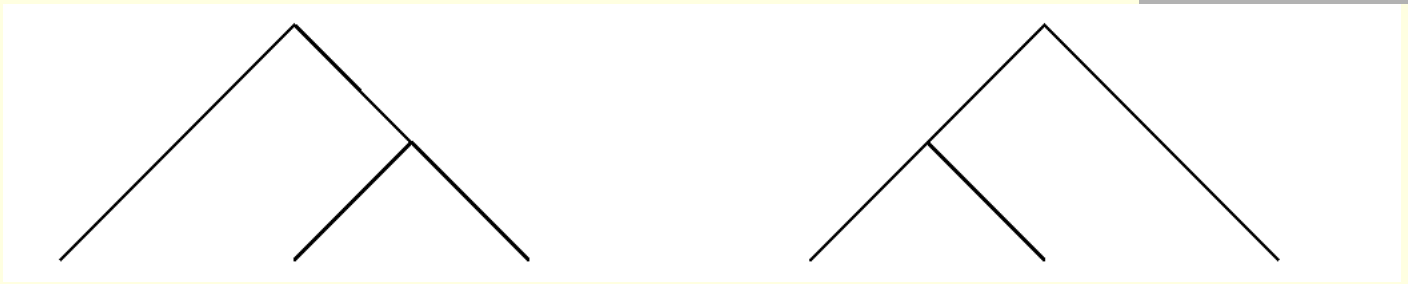
- **New features**
  - paraphrases
  - surface features

- **The ultimate goal**

  Use the Web as a corpus, and not just as a source of page hit frequencies!

Noun Compound Bracketing

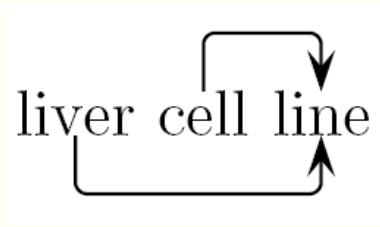
Noun Compound Bracketing: The Problem

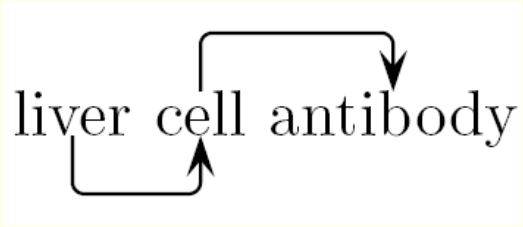


(a) [liver [cell line]] (*right bracketing*)

(b) [[liver cell] antibody] (*left bracketing*)

- In (a), the *cell line* is derived from the *liver*.
- In (b), the *antibody* targets the *liver cell*.

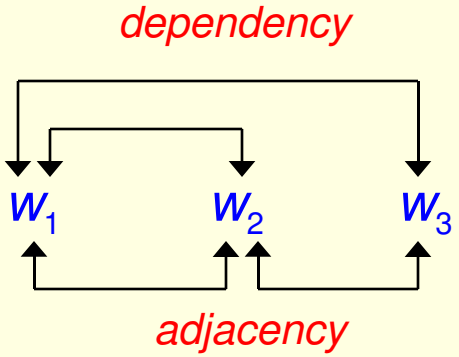




Measuring Word Associations

Using *n*-gram Statistics

- 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

Here starts the new work…

- Observations
  - Authors often disambiguate noun compounds using surface markers.
  - The enormous size of the Web makes them frequent enough to be useful.
- Idea
  - Look for instances of the target noun compound where it occurs with suitable surface markers.

Web-derived Surface Features: Dash (hyphen)

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

Web-derived Surface Features: Possessive Marker

- Attached to the first word
  - *brain**'s** stem cell* ➔ *right*
- Attached to the second word
  - *brain stem**'s** cell* ➔ *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
  - ***b****rain* ***S****tem* ***c****ell* ➔ *right*
  - ***b****rain* ***S****tem* ***C****ell* ➔ *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*
  - *(brain) stem cell* ➔ *right*
- Two words
  - *(growth factor) beta* ➔ *left*
  - *brain (stem cell)* ➔ *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: Dash to External Word

- External word to the left
  - ***mouse-****brain stem cell* ➔ *right*
- External word to the right
  - *tumor necrosis factor-****alpha*** ➔ *left*

Web-derived Surface Features: Problems & Solutions

- **Problem:** search engines ignore punctuation
  - *"brain-stem cell" does not work*

- **Solution:**
  - query for *"brain stem cell"*
  - obtain 1,000 document summaries
  - scan for the features in these summaries

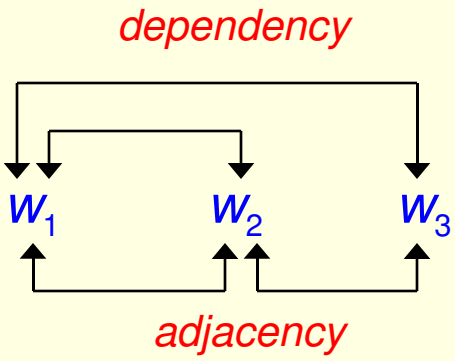
One can get much more than 1,000 results using the "*" operator and inflections.

Other Web-derived Features: Abbreviation

- After the second word
  - *tumor necrosis (TN) factor* ➔ *left*
- After the third word
  - *tumor necrosis factor (NF)* ➔ *right*
- Query for e.g., "*tumor necrosis tn factor*"
  "*tumor necrosis factor nf*"

Other Web-derived Features: Concatenation

- Consider "*health care reform*"
  - *healthcare* : 79,500,000
  - *carereform* : 269
  - *healthreform*: 812
- Adjacency model
  - *healthcare vs. carereform*
- Dependency model
  - *healthcare vs. healthreform*
- Triples
  - *"healthcare reform" vs. "health carereform"*



Tests for lexicalization

Other Web-derived Features: Using the star operator "*"

- Single star
  - "*health care * reform*" ➔ *left*
  - "*health * care reform*" ➔ *right*
- More stars and/or reverse order
  - "*care reform * * health*" ➔ *right*
  - "*reform * * * health care*" ➔ *left*

Other Web-derived Features: Reorder

- Reorders for "*health care reform*"
  - "*care reform health*" ➔ *right*
  - "*reform health care*" ➔ *left*

Other Web-derived Features:
Internal Inflection Variability

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

  ➔ *right*

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

  ➔ *left*

Other Web-derived 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-bracketed?

- **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

- Method: Exact phrase queries limited to English
- Dataset: Lauer's Dataset
  - 244 noun compounds from Grolier's encyclopedia

Noun Compound			Bracketing
minority	business	development	left
satellite	data	systems	right
disaster	relief	assistance	left
county	extension	agents	right
world	food	production	right
granary	storage	baskets	right
customs	enforcement	vehicles	left
airport	security	improvements	left

Evaluation Results (1)

- Co-occurrences

Model	Correct	Wrong	N/A	Prec.	Cover.
# adjacency	183	61	0	75.00±5.79	100.00
Pr adjacency	180	64	0	73.77±5.86	100.00
PMI adjacency	182	62	0	74.59±5.81	100.00
χ^2 adjacency	184	60	0	**75.41±5.77**	100.00
# dependency	193	50	1	79.42±5.52	99.59
Pr dependency (= PMI dep.)	194	50	0	79.51±5.50	100.00
χ^2 dependency	195	49	0	**79.92±5.47**	100.00
# adjacency (*)	152	41	51	78.76±6.30	79.10
# adjacency (**)	162	43	39	79.02±6.08	84.02
# adjacency (***)	150	51	43	74.63±6.44	82.38
# adjacency (*, rev.)	163	48	33	77.25±6.11	86.47
# adjacency (**, rev.)	165	51	28	76.39±6.09	88.52
# adjacency (***, rev.)	156	57	31	73.24±6.32	87.30

Evaluation Results (2)

- Paraphrases, surface features, majority vote

Model	Correct	Wrong	N/A	Prec.	Cover.
Concatenation adjacency	175	48	21	78.48±5.85	91.39
Concatenation dependency	167	41	36	**80.29±5.93**	85.25
Concatenation triples	76	3	165	**96.20±6.78**	32.38
Inflection variability	69	36	139	65.71±9.49	43.03
Swap first two words	66	38	140	63.46±9.58	42.62
Reorder	112	40	92	73.68±7.52	62.30
Abbreviations	21	3	220	**87.50±18.50**	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
Majority vote	210	22	12	90.52±4.46	95.08
Majority vote, then default to 'left'	*218*	*26*	*0*	***89.34±4.50***	***100.00***
Baseline (choose left)	163	81	0	66.80±6.13	100.00

Comparison to Others

Model	Accuracy (%)
baseline (*left*)	66.80±6.13
[Lauer, 1995] adjacency	68.90±6.07
[Lauer, 1995] dependency	77.50±5.65
My χ^2 dependency	*79.92±5.47*
[Lauer, 1995] tuned	80.70±5.41
My majority vote → left	***89.34±4.50***
[Lapata and Keller, 2004]: baseline (122 examples)	63.93±8.83
[Lapata and Keller, 2004]: best *BNC* (122 examples)	68.03±8.72
[Lapata and Keller, 2004]: best *AltaVista* (122 examples)	78.68±8.08
*[Girju et al., 2005]: best C5.0 (shuffled dataset)	83.10±5.20

Application: Query Segmentation

S. Bergsma, Q. Wang.
Learning Noun Phrase Query Segmentation.
EMNLP'07, pp. 819-826.

- Segmentation
  - [used car parts]
  - [used car] [parts]
  - [used] [car parts]
  - [used] [car] [parts]
- Bracketing
  - [[used car] parts]
  - [used [car parts]]

ACL'07: Adding Noun Phrase Structure to the Penn Treebank
David Vadas and James R. Curran

Prepositional Phrase Attachment

PP attachment

PP combines with the NP
to form another NP

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

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

PP is an indirect
object of the verb

quadruple: (v, n1, p, n2)

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

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

Results

Model	P(%)	R(%)
Baseline (noun attach)	41.82	100.00
$\#(x, p)$	58.91	83.97
$\Pr(p\|x)$	66.81	83.97
$\Pr(p\|x)$ smoothed	**66.81**	83.97
$\#(x, p, n_2)$	65.78	81.02
$\Pr(p, n_2\|x)$	68.34	81.62
$\Pr(p, n_2\|x)$ smoothed	**68.46**	83.97
(1) "$v\ n_2\ n_1$"	**59.29**	22.06
(2) "$p\ n_2\ v\ n_1$"	**57.79**	71.58
(3) "n_1 * $p\ n_2\ v$"	**65.78**	20.73
(4) "$v\ p\ n_2\ n_1$"	**81.05**	8.75
(5) "v *pronoun* $p\ n_2$"	**75.30**	30.40
(6) "*be* $n_1\ p\ n_2$"	**63.65**	30.54
n_1 is *pronoun*	**98.48**	3.04
v is *to be*	**79.23**	9.53
Surface features (summed)	**73.13**	9.26
Maj. vote, of $\rightarrow$ *noun*	85.01±1.21	91.77
Maj. vote, of $\rightarrow$ *noun, N/A* $\rightarrow$ *verb*	***83.63±1.30***	***100.00***

Simpler but not significantly different from 84.3% (Pantel&Lin,00).

Noun Phrase Coordination

NP Coordination: Ellipsis

- Ellipsis
  - *car and truck production*
    - means *car production and truck production*

- No ellipsis
  - *president and chief executive*

NP Coordination: Ellipsis

- Penn Treebank annotations
  - ellipsis:

    (NP *car/NN and/CC truck/NN production/NN****)***.
  - no ellipsis:

    (NP

    (NP *president/NN****)***

    and/CC

    (NP *chief/NN executive/NN****))***

Results
428 examples from Penn TB

Model	P(%)	R(%)
Baseline: ellipsis	56.54	100.00
(n_1, h) vs. (n_2, h)	**80.33**	28.50
(n_1, h) vs. (n_1, c, n_2)	61.14	45.09
(n_2, c, n_1, h)	**88.33**	14.02
(n_2, h, c, n_1)	**76.60**	21.96
(n_1, h, c, n_2, h)	**75.00**	6.54
(n_2, h, c, n_1, h)	**78.67**	17.52
Heuristic 1	**75.00**	0.93
Heuristic 4	64.29	6.54
Heuristic 5	61.54	12.15
Heuristic 6	**87.09**	7.24
Number agreement	**72.22**	46.26
Surface sum	**82.80**	21.73
Majority vote	83.82	80.84
Majority vote, N/A → no ellipsis	***80.61***	***100.00***

Comparable to other researchers (but no standard dataset).

Paraphrasing Noun Compounds

Noun Compound Semantics

- **Traditionally – 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
  - **Recoverably Deletable Predicates** (Levi,1978):
    - *malaria mosquito* ➔ CAUSE
    - *olive oil* ➔ FROM
- **Our 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

NC Semantics: Method

pre-modifier

- For NC "**noun1 noun2**", query for:

"noun2 THAT * noun1"

post-modifier

THAT can be *that*, *which* or *who; up to 8 "*"s.*

- POS tag the snippets.
- Extract verbal paraphrases.

NC Semantics: Sample Verbal Paraphrases

Verbs+prepositions for *migraine treatment*

7	prevent
3	be given for
3	be for
2	reduce
2	benefit
1	relieve

Example: Treatments

- Dynamic componential analysis

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

- Classic componential analysis

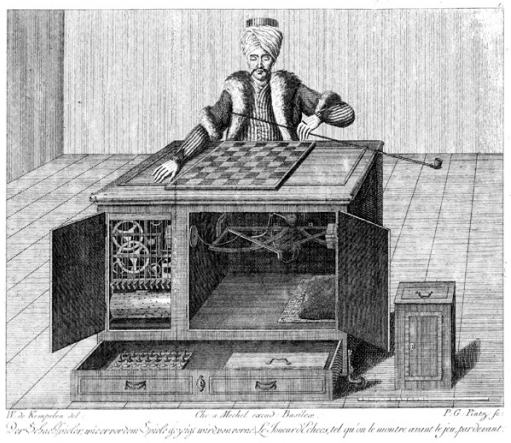
	man	woman	boy	bull
ANIMATE	+	+	+	+
HUMAN	+	+	+	−
MALE	+	−	+	+
ADULT	+	+	−	+

Comparing to (Girju&al.,05)

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)~~

14 out of 21 relations are shown.

Amazon's Mechanical Turk: *Malaria Mosquito*



- **Five judges:**
  - **5 carries**
  - **3 causes**
  - **2 transmits**
  - **2 infects with**
  - **1 has**
  - 1 supplies
- **The program:**
  - **23 carry**
  - 16 spread
  - **12 cause**
  - **9 transmit**
  - 7 bring
  - **4 have**
  - **3 be infected with**
  - 3 be responsible for
  - …

MTurk: Comparison to 30 Humans

0.96 "blood donor" NOMINALIZATION:AGENT

MTurk: give(30), donate(16), supply(8), provide(6), share(2), contribute(1), volunteer(1), offer(1), choose(1), hand over(1), ...

Web: give(653), donate(395), receive(74), sell(41), provide(39), supply(17), be(13), match(11), contribute(10), offer(9), ...

0.95 "women professors" BE

MTurk: be(22), teach(2), look like(2), be born(2), research(1), study(1), be comprised of(1), behave like(1), include(1), be gendered(1), ...

Web: be(251), teach(46), study(38), specialize in(26), appear as(13), think(10), take(10), research(9), work with(9), think that(9), ...

0.94 "student friends" BE

MTurk: be(18), come from(2), help(2), be had by(2), be made by(2), include(1), involve(1), act like(1), live as(1), work as(1), ...

Web: be(1250), have(34), support(33), know(31), teach(22), give(21), meet as(19), help(16), guide(15), pose as(12), ...

0.93 "city wall" $HAVE_2$

MTurk: surround(24), protect(10), enclose(8), encircle(7), encompass(3), be in(3), contain(2), snake around(1), border(1), go around(1), ...

Web: surround(708), encircle(203), protect(191), divide(176), enclose(72), separate(49), ring(41), be(34), encompass(25), defend(25), ...

Average cosine correlation

- Average cosine correlation (in %) between human- and program-generated verbs for the *Levi-250 dataset*.

Min # of Web Verbs	Number of Compounds	Correlation with Humans Using All Verbs	Correlation with Humans First Verb Only
0	250	31.8%	30.6%
1	236	33.7%	32.4%
3	216	35.4%	34.1%
5	203	36.9%	35.6%
10	175	37.3%	35.5%

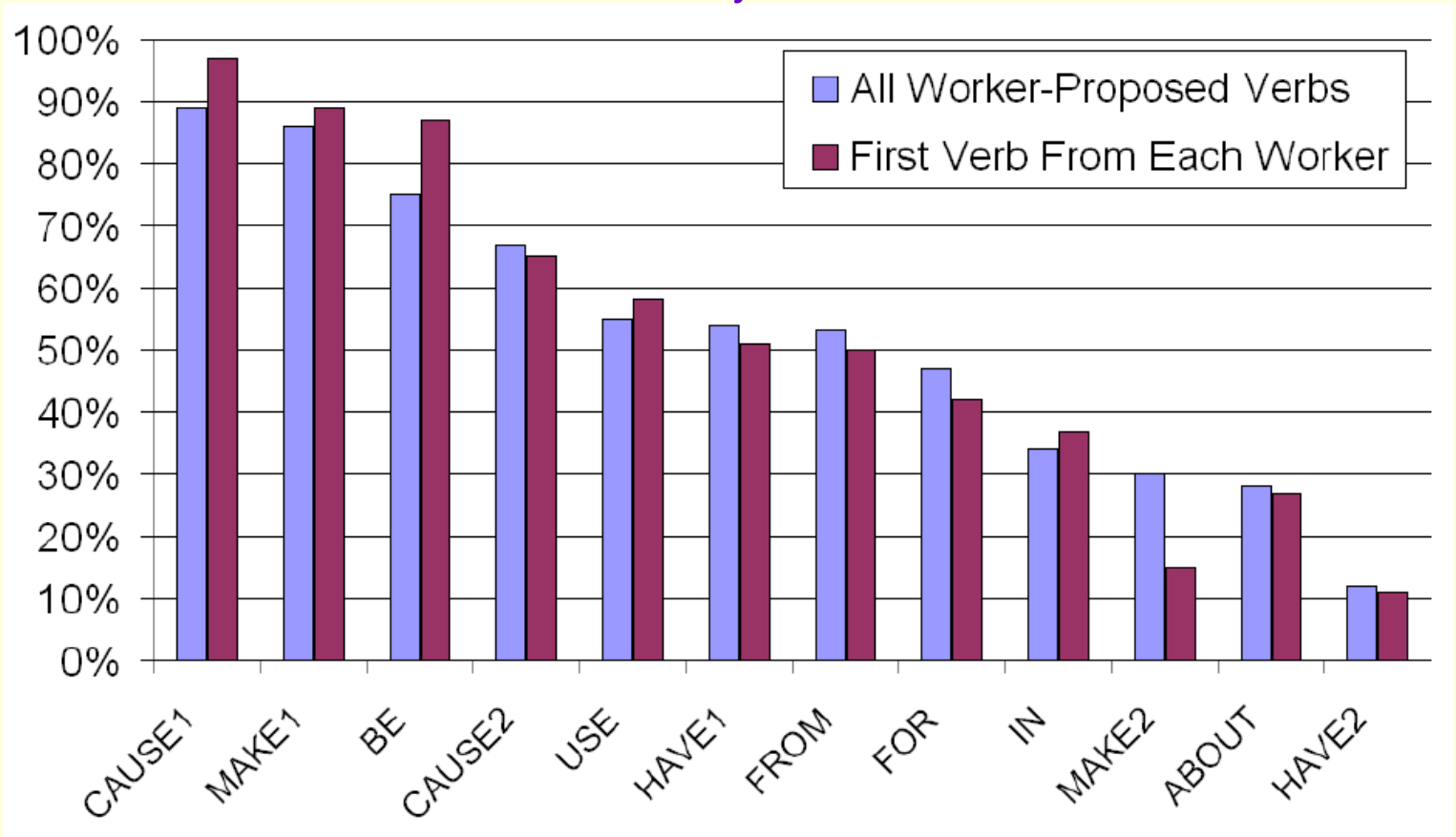
$$cos(\overrightarrow{h}, \overrightarrow{p}) = \frac{\sum_{i=1}^{n} h_i p_i}{\sqrt{\sum_{i=1}^{n} h_i^2} \sqrt{\sum_{i=1}^{n} p_i^2}}$$

Levi's Recoverably Deletable Predicates

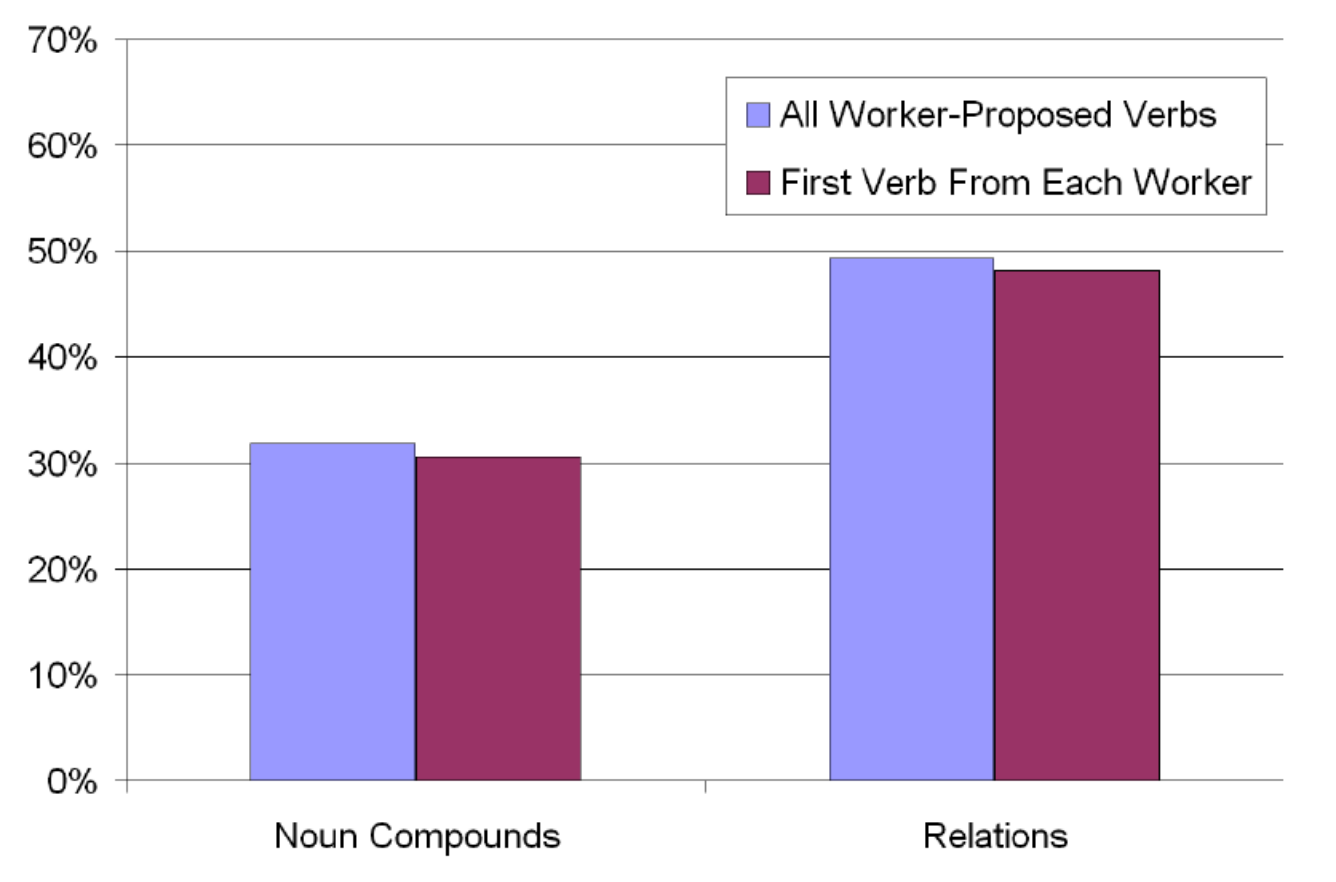
RDP	Example	Subj/obj	Traditional Name
CAUSE_1	*tear gas*	object	causative
CAUSE_2	*drug deaths*	subject	causative
HAVE_1	*apple cake*	object	possessive/dative
HAVE_2	*lemon peel*	subject	possessive/dative
MAKE_1	*silkworm*	object	productive/composit.
MAKE_2	*snowball*	subject	productive/composit.
USE	*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

MTurk (human) vs. Web (program): Aggregated by Levi's RDP

- Cosine correlation (in %s) between the human- and the program-generated verbs by Levi's RDP: using all human-proposed verbs vs. using the first verb from each worker only.



Average Cosine Correlation



- Left: calculated for each noun compound
- Right: aggregated by relation

Predicting Abstract Semantic Relations

Levi's Recoverably Deletable Predicates

RDP	Example	Subj/obj	Traditional Name
CAUSE_1	*tear gas*	object	causative
CAUSE_2	*drug deaths*	subject	causative
HAVE_1	*apple cake*	object	possessive/dative
HAVE_2	*lemon peel*	subject	possessive/dative
MAKE_1	*silkworm*	object	productive/composit.
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

Search Engine Queries

- Given *noun1* and *noun2*, query for:

  “noun2 * noun1"
  “noun1 * noun2"

  Use up to 8 “*”s.

- POS tag the snippets.
- Extract: verbs, prep, verb+prep, coordinations.

Most Frequent Features for *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

Predicting Semantic Relations: Levi's RDPs

Model	Accuracy	Coverage	Avg #feats	Avg Σfeats
Human: all v	78.4±6.0	99.5	34.3	70.9
Human: first v from each worker	72.3±6.4	99.5	11.6	25.5
Web: $v + p + c$	50.0±6.7	99.1	216.6	1716.0
Web: $v + p$	50.0±6.7	99.1	208.9	1427.9
Web: $v + c$	46.7±6.6	99.1	187.8	1107.2
Web: v	45.8±6.6	99.1	180.0	819.1
Web: p	33.0±6.0	99.1	28.9	608.8
Web: $p + c$	32.1±5.9	99.1	36.6	896.9
Baseline (majority class)	19.6±4.8	100.0	–	–

- **Vector-space model**
- **kNN Classifier**
- **Dice coefficient (freqs)**

- *v* – verb
- *p* – preposition
- *c* – coordinating conjunction

Relations Between Complex Nominals

SemEval'07: Data

"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"

There are seven such relations.

SemEval'07: Results

Team	P	R	F	Acc
A – WordNet = NO & Query = NO				
UCD-FC	66.1	66.7	64.8	66.0
ILK	60.5	69.5	63.8	63.5
UCB	62.7	63.0	62.7	65.4
UMELB-B	61.5	55.7	57.8	62.7
UTH	56.1	57.1	55.9	58.8
UC3M	48.2	40.3	43.1	49.9
avg±stdev	58.7±6.9	59.3±11.6	58.1±9.1	60.7±6.7

kNN classifier with the Dice coefficient

Using up to 10 stars: 67.0

SemEval'07: Results

Team	P	R	F	Acc
C – WordNet = NO & Query = YES				
UCB	64.2	66.5	65.1	67.0
UCD-FC	66.1	66.7	64.8	66.0
UC3M	49.4	43.9	45.3	50.1
avg±stdev	59.9±9.1	59.0±13.1	58.4±11.3	61.0±9.5

kNN classifier with the Dice coefficient

Using up to 10 stars: 68.1

SAT Analogy Questions

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

Table 2: **SAT analogy examples from the set of 374.** The stems are in **bold**, the solutions are in *italic*, and the distractors are in plain text.

SAT: Nouns Only

Model	✓	×	∅	Accuracy	Cover.
$v + p + c$	129	52	3	**71.3±7.0**	**98.4**
v	122	56	6	68.5±7.2	96.7
$v + p$	119	61	4	66.1±7.2	97.8
$v + c$	117	62	5	65.4±7.2	97.3
$p + c$	90	90	4	50.0±7.2	97.8
p	84	94	6	47.2±7.2	96.7
baseline	37	147	0	20.0±5.2	100.0
LRA	122	59	3	67.4±7.1	98.4

Head-Modifier Relations in Noun-Noun Compounds

30-Relations from (Nastase & Szpakowicz,2003)

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	cont	apple cake	M is contained in H
equative	eq	player coach	H is also M
material	mat	brick house	H is made of M
measure	meas	expensive book	M is a measure of H
topic	top	weather report	H is concerned with M
type	type	oak tree	M is a type of H

Noun-Modifier Relations: 30 classes

Model	✓	×	∅	Accuracy	Cover.
$v + p$	240	352	8	**40.5±3.9**	**98.7**
$v + p + c$	238	354	8	40.2±3.9	98.7
v	234	350	16	40.1±3.9	97.3
$v + c$	230	362	8	38.9±3.8	98.7
$p + c$	114	471	15	19.5±3.0	97.5
p	110	475	15	19.1±3.0	97.5
baseline	49	551	0	8.2±1.9	100.0
LRA	239	361	0	39.8±3.8	100.0

- *v* – verb
- *p* – preposition
- *c* – coordinating conjunction

Application to Machine Translation

MT: 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 an
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

Paraphrasing the Phrase Table (1)

- **Phrase Table Entry**
  - , spain 's economy ||| , la economía española ||| 1 0.0056263 1 0.00477047 2.718
- **Paraphrased Entries**
  - , economy of spain ||| , la economía española ||| 1 0.0056263 1 0.00477047 2.718
  - , the economy of spain ||| , la economía española ||| 1 0.0056263 1 0.00477047 2.718
  - , spain economy ||| , la economía española ||| 1 0.0056263 1 0.00477047 2.718
  - , economy of a spain ||| , la economía española ||| 1 0.0056263 1 0.00477047 2.718
  - , economy of an spain ||| , la economía española ||| 1 0.0056263 1 0.00477047 2.718
  - , economy of the spain ||| , la economía española ||| 1 0.0056263 1 0.00477047 2.718

Web-based filtering

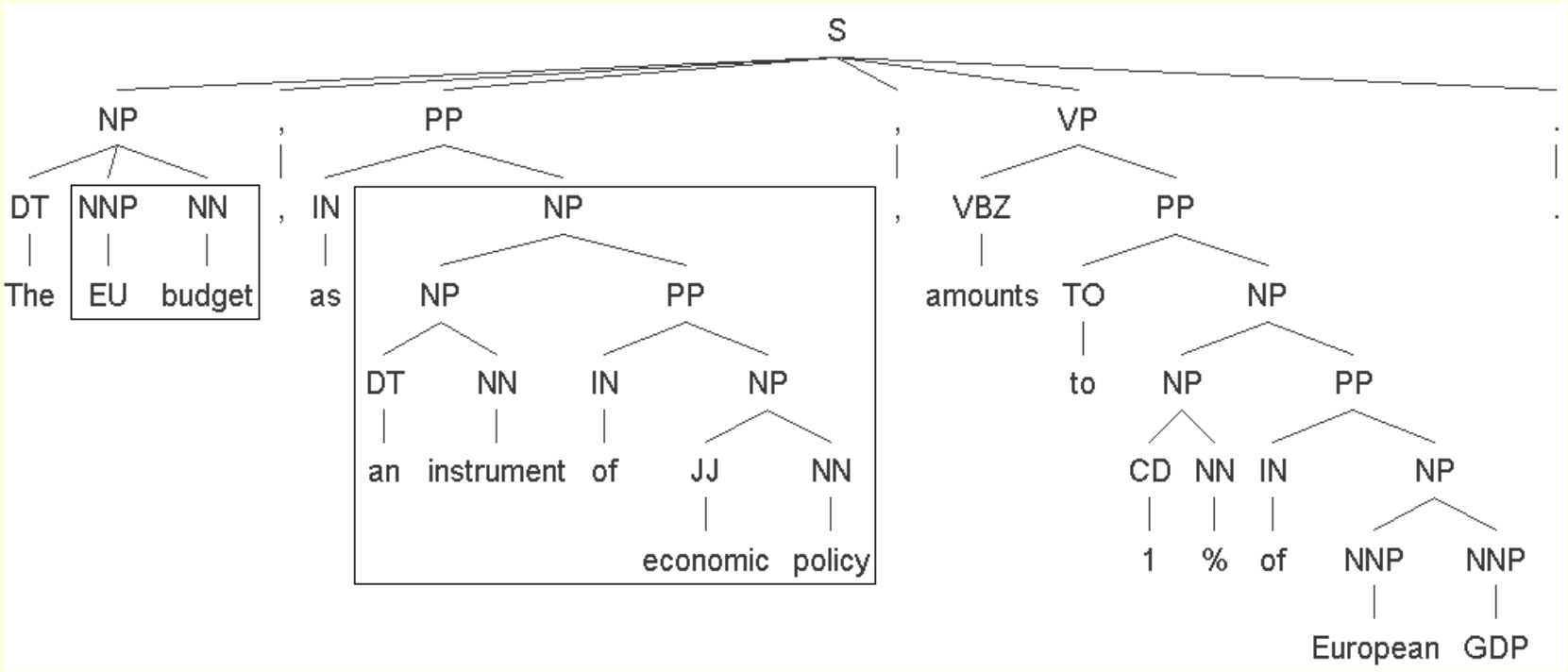
Paraphrasing the Phrase Table (2)

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 the Training Corpus

We must cooperate internationally , and this should include UN initiatives .
We must cooperate internationally , and this should include *initiatives of the UN* .
We must cooperate internationally , and this should include *initiatives at the UN* .
We must cooperate internationally , and this should include *initiatives in the UN* .
Both reports on economic policy confirm the impression that environment policy is only a stepchild .
Both reports on economic policy confirm the impression that *policy on the environment* is only a stepchild .
Both reports on economic policy confirm the impression that *policy on environment* is only a stepchild .
Both reports on economic policy confirm the impression that *policy for the environment* is only a stepchild .
Both economic policy reports confirm the impression that environment policy is only a stepchild .
Both economic policy reports confirm the impression that *policy on the environment* is only a stepchild .
Both economic policy reports confirm the impression that *policy on environment* is only a stepchild .
Both economic policy reports confirm the impression that *policy for the environment* is only a stepchild .
To the contrary , what is needed now - absolutely needed - is a reduction in the taxation of labour .
To the contrary , what is needed now - absolutely needed - is a reduction in the labour taxation .
To the contrary , what is needed now - absolutely needed - is a labour taxation reduction .
To the contrary , what is needed now - absolutely needed - is a reduction in the labour 's taxation .
To the contrary , what is needed now - absolutely needed - is a labour 's taxation reduction .

Paraphrasing a Sentence



Paraphrasing NPs/NCs

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 stats

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.

Results

System	Bleu score 10k	20k	40k	80k
S (baseline)	22.38	24.33	26.48	27.05
S_{parW}	22.57	24.41	25.96	
$S^{\star}$	22.58	25.00	26.48	
$S + S_{parW}$	**23.05**	**25.01**	**26.75**	

S_{parW} original corpus, augmented with sentence-level paraphrases, all transformations;

$\star$ means paraphrasing the phrase table;

$+$ means merging the phrase tables;

Conclusion

Conclusion

- Tapped the potential of very large corpora for unsupervised algorithms:
  - Go beyond n-grams
    - Surface features
    - Paraphrases
  - Results
    - competitive with the best unsupervised algorithms
    - can rival supervised algorithms

Resume

- Surface Features & Paraphrases
- Syntactic Tasks
  - Noun Compound Bracketing
  - Prepositional Phrase Attachment
  - Noun Compound Coordination
- Semantic Tasks
  - Paraphrasing Noun Compounds
  - Predicting Abstract Semantic Relations
  - Relations Between Complex Nominals
  - SAT Analogy Questions
  - Head-Modifier Relations
- Application
  - Machine Translation

Future Work

- New exciting features
- Other problems
- Use less queries
- **Use the Web as a corpus, and not just as a source of page hit frequencies!**

Thank You

Questions?