Microsoft Research

Microsoft Research Asia
Faculty Summit 2012

Deep Natural Language Processing for Improving a Search Engine using Cloud Computing

Daisuke Kawahara Sadao Kurohashi

Kyoto University

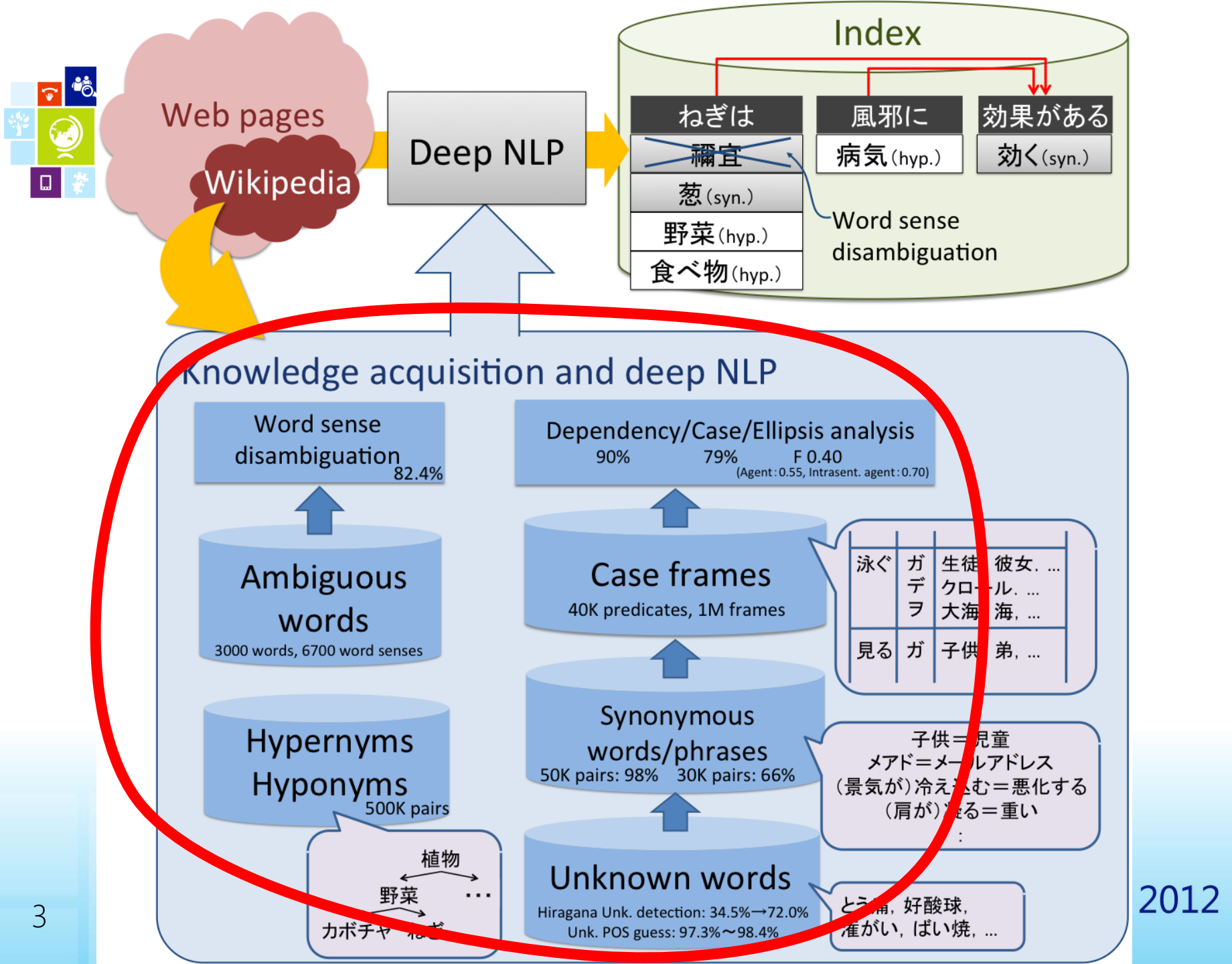
Microsoft Research Asia
Faculty Summit 2012

Overview

Web services of search and analysis based on deep NLP

Deep NLP-based dexing and search

- PC cluster/grid
  1,000～5,000 CPUs
- Cloud
  ～10,000～ CPUs


Web pages
Wikipedia
Deep NLP
Index
ねぎは
風邪に
効果がある
禰宜
病気(hyp.)
効く(syn.)
葱(syn.)
野菜(hyp.)
食べ物(hyp.)
Word sense disambiguation
Knowledge acquisition and deep NLP
Word sense disambiguation
82.4%
Dependency/Case/Ellipsis analysis
90%
79%
F 0.40
(Agent:0.55, Intrasent. agent:0.70)
Ambiguous words
3000 words, 6700 word senses
Case frames
40K predicates, 1M frames
泳ぐ
ガ
デ
ヲ
生徒 彼女, ...
クロール, ...
大海 海, ...
見る
ガ
子供 弟, ...
Hypernyms
Hyponyms
500K pairs
Synonymous words/phrases
50K pairs: 98%
30K pairs: 66%
子供＝児童
メアド＝メールアドレス
(景気が)冷え込む＝悪化する
(肩が)凝る＝重い
植物
野菜
...
カボチャ
Unknown words
Hiragana Unk. detection: 34.5%→72.0%
Unk. POS guess: 97.3%～98.4%
好酸球,
ばい焼, ...

2012

ThinkPad – Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/ThinkPad

Google

Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Català
Česky
Dansk
Deutsch
Eesti
Español
Français

ThinkPad

From Wikipedia, the free encyclopedia

The **ThinkPad** is a line of laptop computers and tablets originally designed, developed and sold by IBM but now produced by Lenovo. They are known for their boxy black design, which was modeled after a traditional Japanese *Bento* lunchbox.[1] Lenovo purchased [illegible]ess and acquired the ThinkPad brand in 2005.

Hypernym

ThinkPads are popular with large busi[illegible]del ThinkPads are revered by technology enthusiasts, collectors and power users due to their durable design, relatively high resale value, and abundance of aftermarket replacement parts. The ThinkPad has been used in space, and is the only laptop certified for use on the International Space Station.[2]

Contents [hide]

1 History
- 1.1 Name
- 1.2 Early models
- 1.3 Industrial design
- 1.4 Reviews and awards
- 1.5 Use in space
- 1.6 Acquisition by Lenovo
- 1.7 Manufacturing

2 Recent models
- 2.1 ThinkPad tablets
  - 2.1.1 ThinkPad Tablet 2
  - 2.1.2 ThinkPad Tablet
- 2.2 T Series

ThinkPad series

IBM ThinkPad R51

Developer	IBM (1992–2005) Lenovo (since 2005)
Type	Laptop
Release date	1992
Operating system	Windows

Kno
and

Case frames

泳ぐ swim
- {人 person, 子 child,…}が
- {クロール crawl, 平泳ぎ,…}で
- {海 sea, 大海,…}を

見る see
- {人 person, 者,…}が
- {望遠鏡 telescope, 双眼鏡 ,…}で
- {姿 figure, 人 person,…}を

Mary **ate** the s

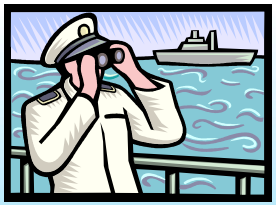
Mary ate **the s**

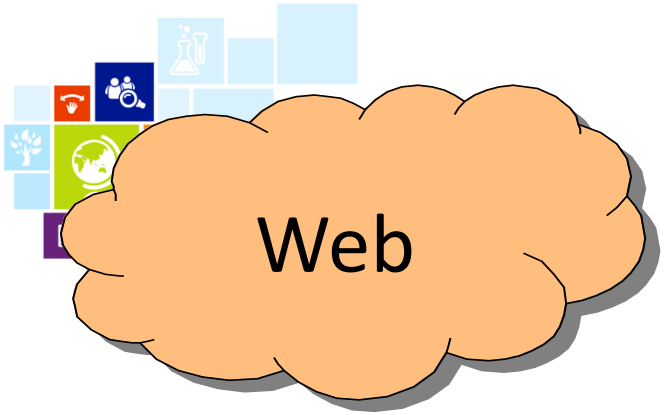
クロールで**泳いでいる**女の子を見た

crawl swim girl saw

望遠鏡で泳いでいる女の子を**見た**

telescope swim girl saw





15G sentences (3G pages)

Parsing and Filtering

89.0% for all
98.3% for 20.7% PAs

2days

Predicate-argument structures

2weeks

Clustering

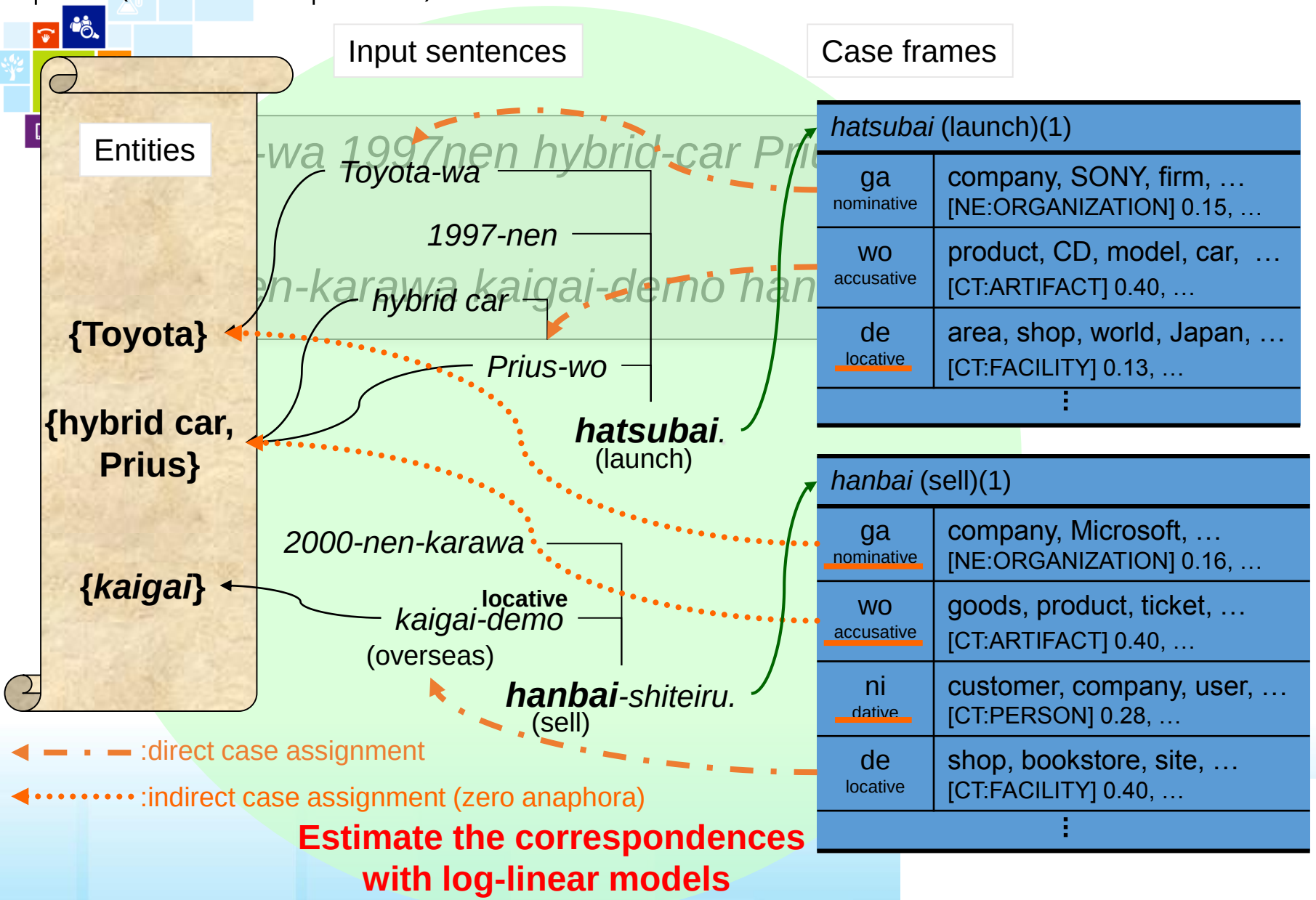
Case frames for 40K predicates

89.0% → 89.7%

Case frame examples

	CS	examples
yaku (1) (bake)	*ga*	I:18, person:15, craftsman:10, …
	wo	bread:2484, meat:1521, cake:1283, …
	de	oven:1630, frying pan:1311, …
yaku (2) (have difficulty)	*ga*	teacher:3, government:3, person:3, …
	wo	hand:2950
	ni	attack:18, action:15, son:15, …
yaku (3) (copy; burn CDR)	*ga*	maker:1, distributor:1, …
	wo	data:178, file:107, copy:9, …
	ni	R:1583, CD:664, CDR:3, …
…		

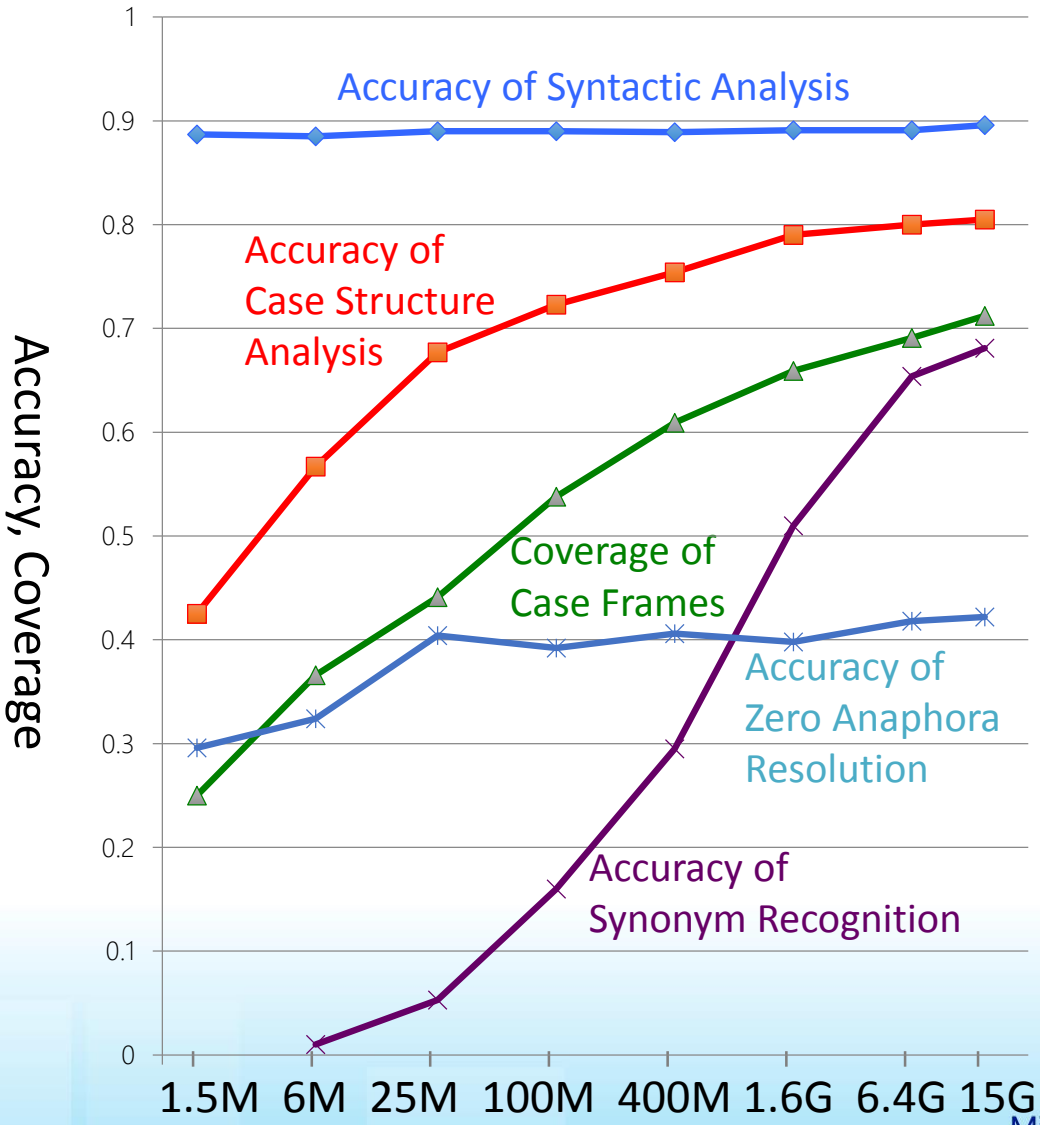
Ellipsis (Zero Anaphora) Resolution

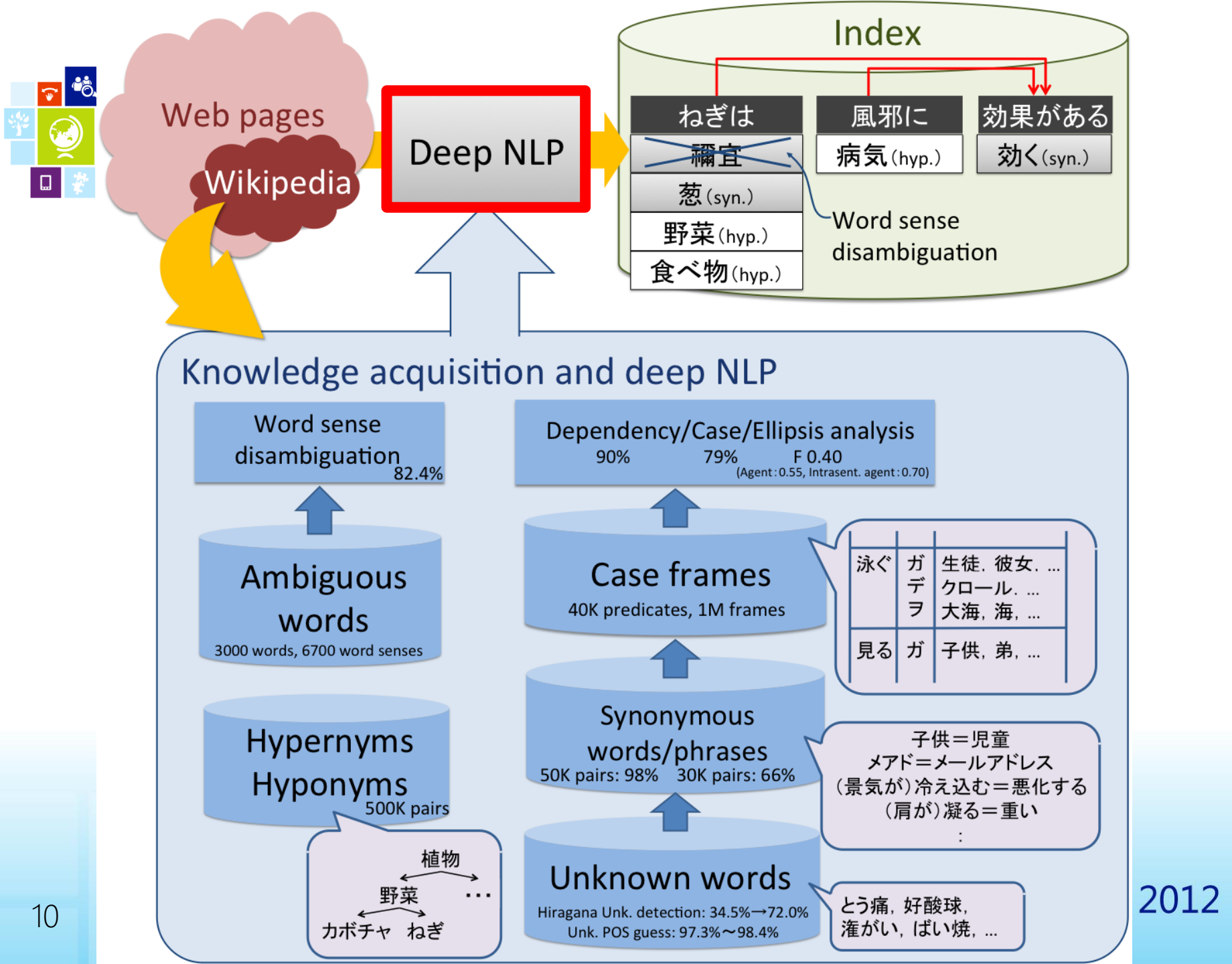


Estimate the correspondences with log-linear models

Toyota launched the hybrid car Prius in 1997. Φ_1 started selling Φ_2 overseas in 2000.

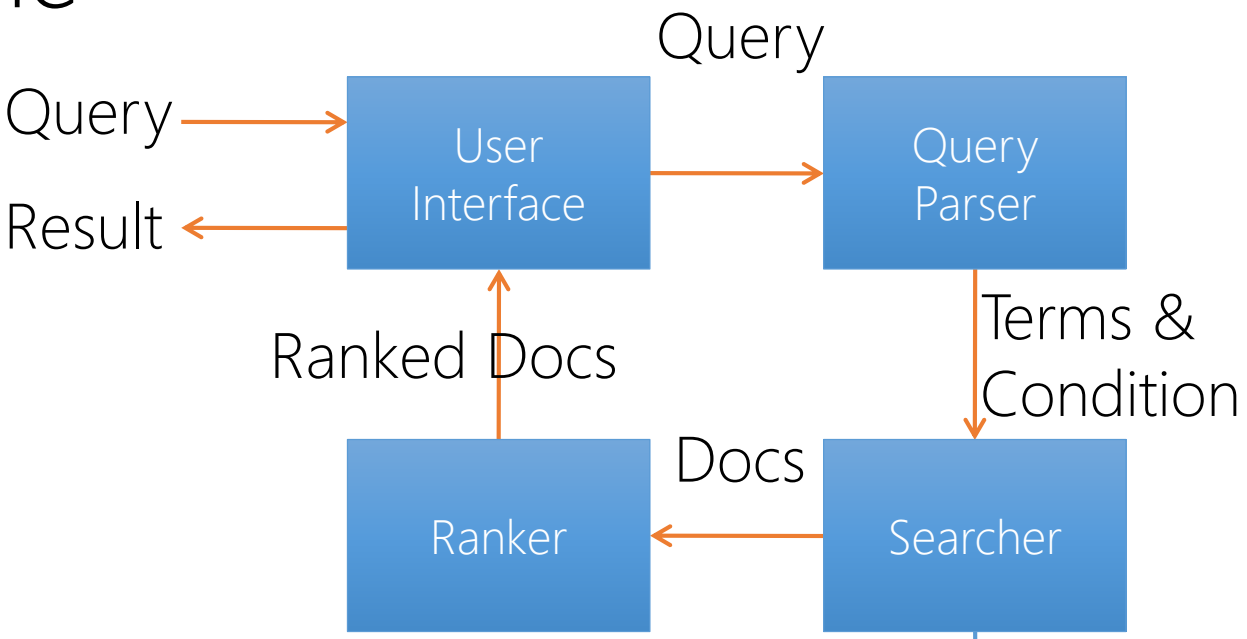
Learning Curve



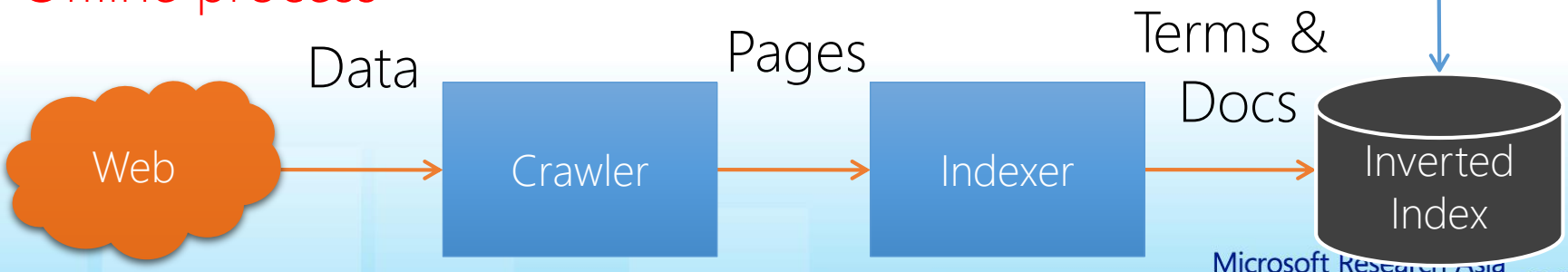

Web pages
Wikipedia
Deep NLP
Index
ねぎは
風邪に
効果がある
禰宜
病気(hyp.)
効く(syn.)
葱(syn.)
野菜(hyp.)
食べ物(hyp.)
Word sense disambiguation
Knowledge acquisition and deep NLP
Word sense disambiguation
82.4%
Ambiguous words
3000 words, 6700 word senses
Hypernyms
Hyponyms
500K pairs
植物
野菜
…
カボチャ ねぎ
Dependency/Case/Ellipsis analysis
90% 79% F 0.40
(Agent:0.55, Intrasent. agent:0.70)
Case frames
40K predicates, 1M frames
泳ぐ ガ 生徒，彼女，…
デ クロール，…
ヲ 大海，海，…
見る ガ 子供，弟，…
Synonymous words/phrases
50K pairs: 98% 30K pairs: 66%
子供＝児童
メアド＝メールアドレス
(景気が)冷え込む＝悪化する
(肩が)凝る＝重い
:
Unknown words
Hiragana Unk. detection: 34.5%→72.0%
Unk. POS guess: 97.3%～98.4%
とう痛，好酸球，
潅がい，ばい焼，…

Architecture of Search Engine

Online process

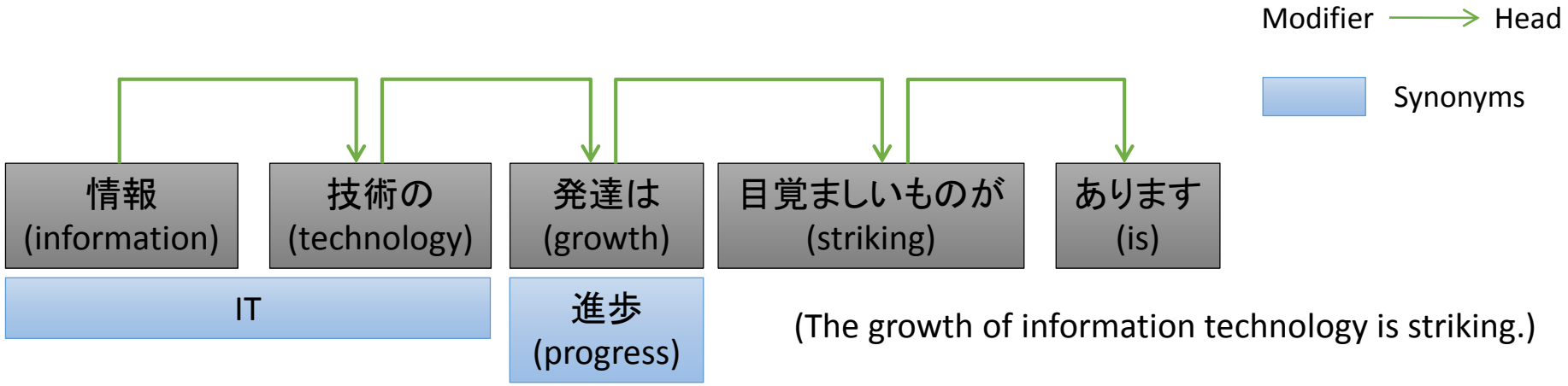


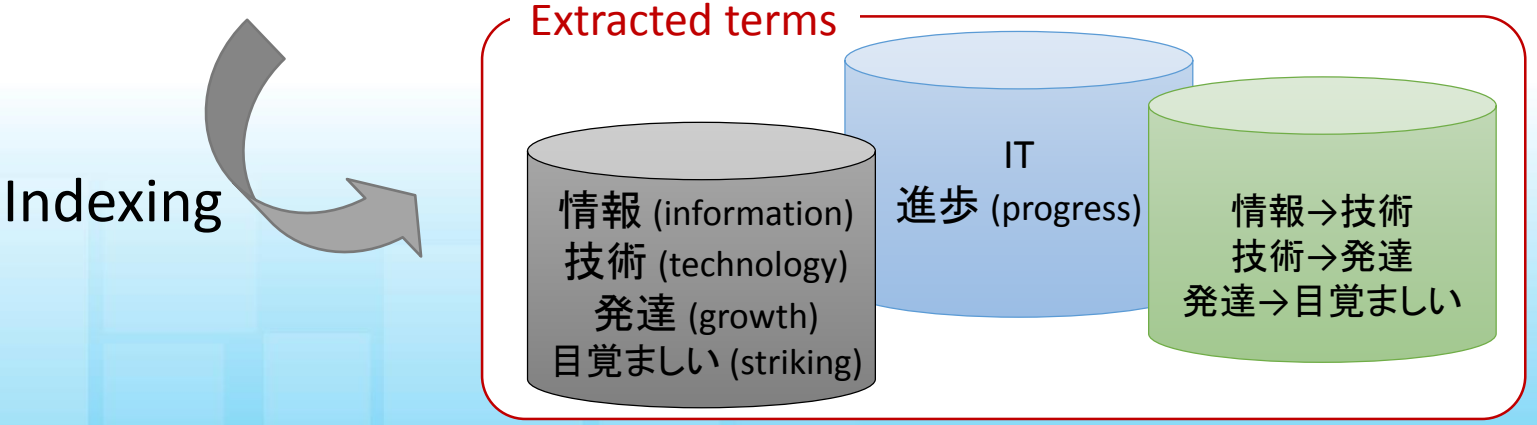
Offline process



Deep NLP-based Indexing

Terms: Words, synonyms/hypernyms of words and dependency relations including zero anaphora



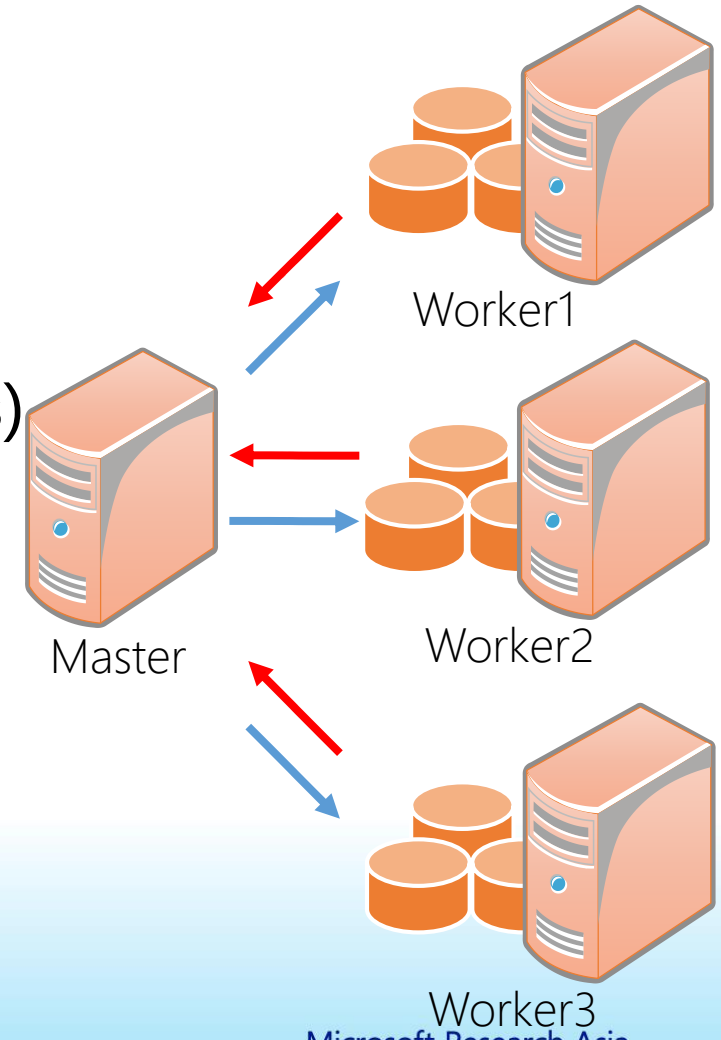


Processing using HPC / Cloud

- Tasks
  - Case frame Compilation from 15G sentences
    ～ 200,000 CPU hours
  - Deep NLP Indexing of 100M Web pages
    ～ 1,500,000 CPU hours
- Using
  - HPC (PC cluster / grid)
    - TITECH TSUBAME 2.0 (～5,000 CPU cores)
    - Kyoto Univ. Super Computer (～3,000 CPU cores)
    - Info-plosion InTrigger (～1,000 CPU cores)
  - Cloud
    - Windows Azure (～10,000～ CPU cores)

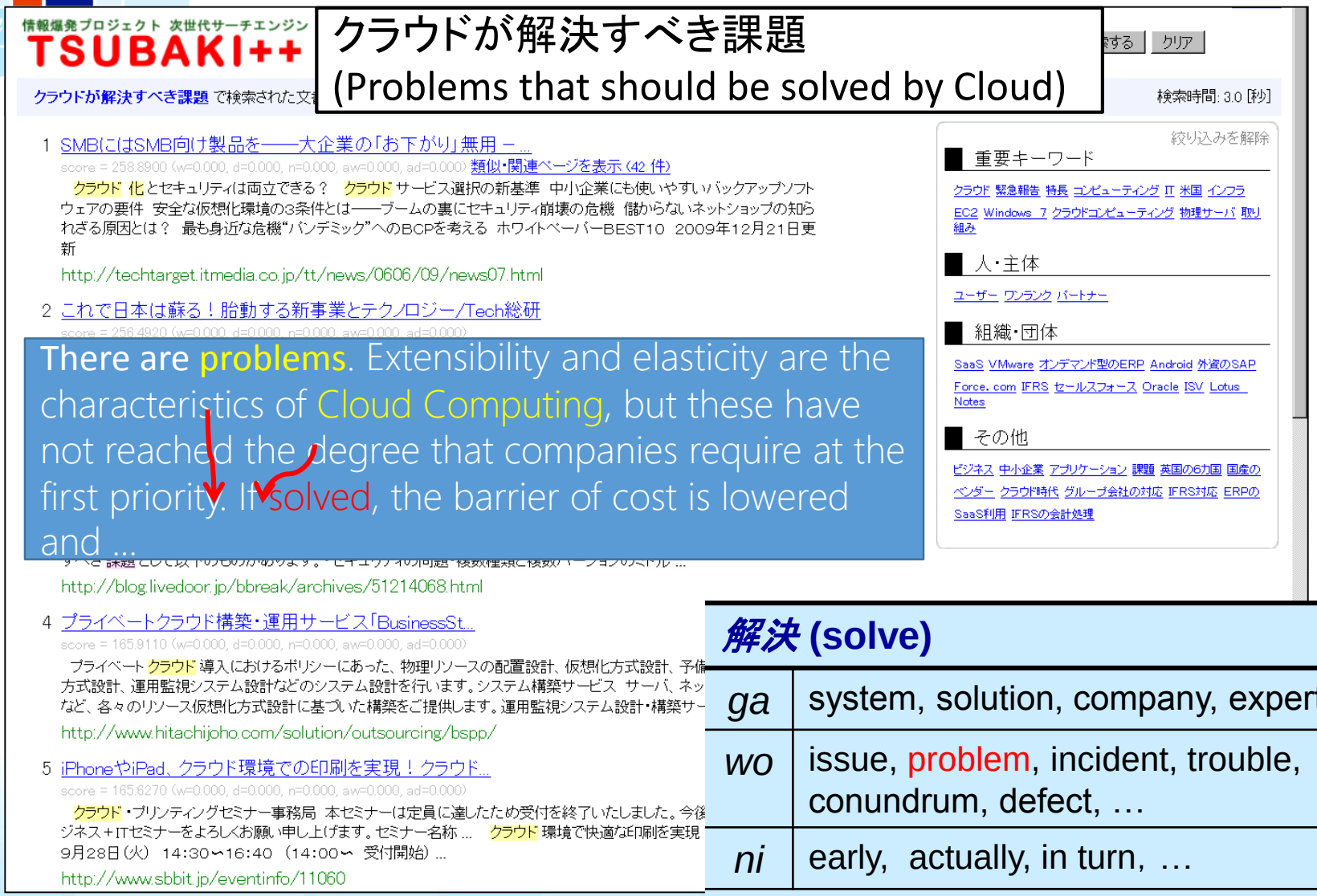
Our Model of Deployment

- Task is completely independent
  - Each task processes a small set of Web pages (100 pages)
- Master/Worker model
  - Master (Web Role)
  - Worker (Worker Role)



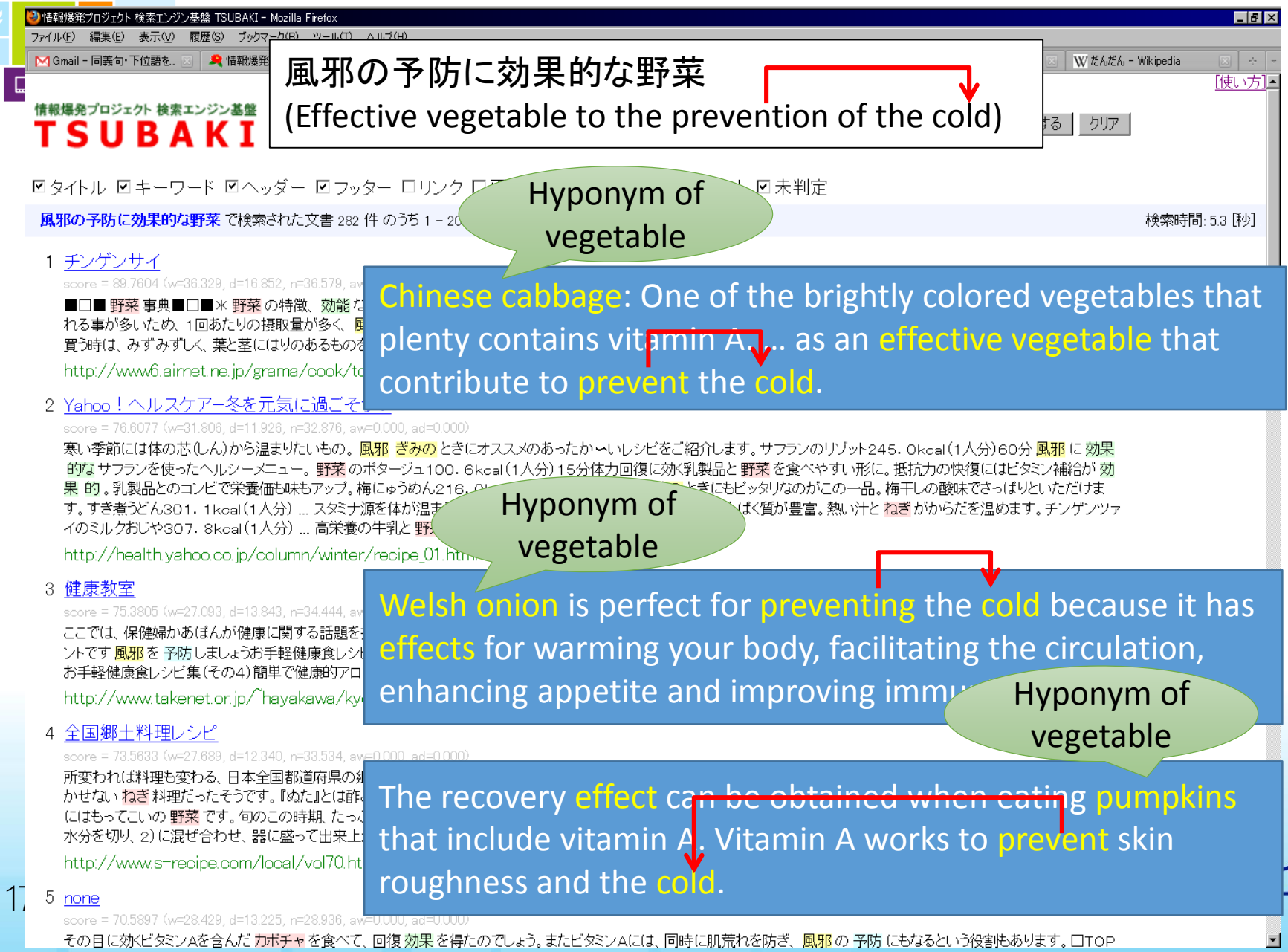
Search Example: Use of Dependency Structure

Search Example: Use of P-A Structure



***解決* (solve)**	
ga	system, solution, company, expert, …
wo	issue, problem, incident, trouble, conundrum, defect, …
ni	early, actually, in turn, …
	:

Search Example: Hypernym/Hyponym Relations



Comparison of Search Engines

- Test set: NTCIR test collection (11M Japanese web pages; 127 natural language queries)

Search engine	MAP	P@3	P@10	nDCG @10
Apache Solr 3.6.1	0.124	0.297	0.276	0.166
TSUBAKI	0.173	0.442	0.379	0.237

Average precision

Precision of top 3 docs

Precision of top 10 docs

Precision of top 10 docs considering rank and relevancy

Web Services on TSUBAKI

Information Analysis with WISDOM

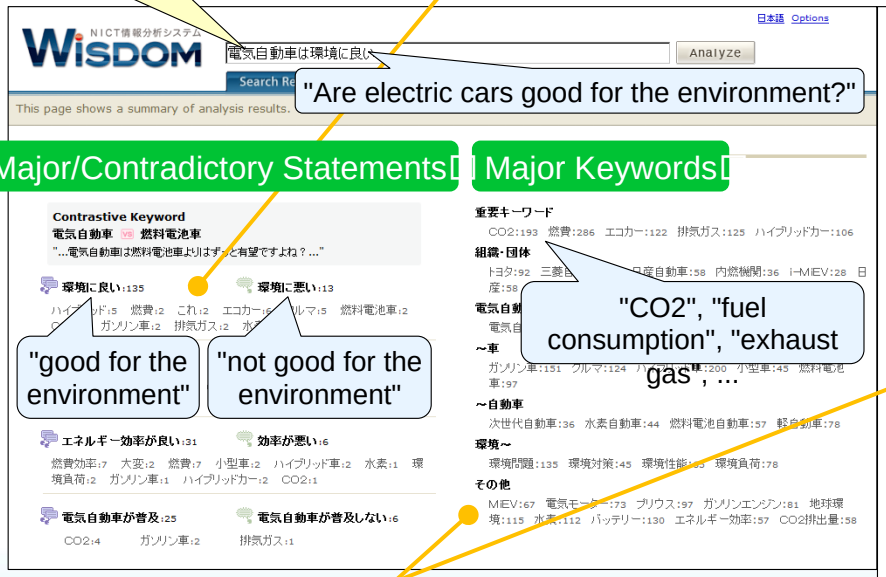
Analyzing web pages based on various criteria

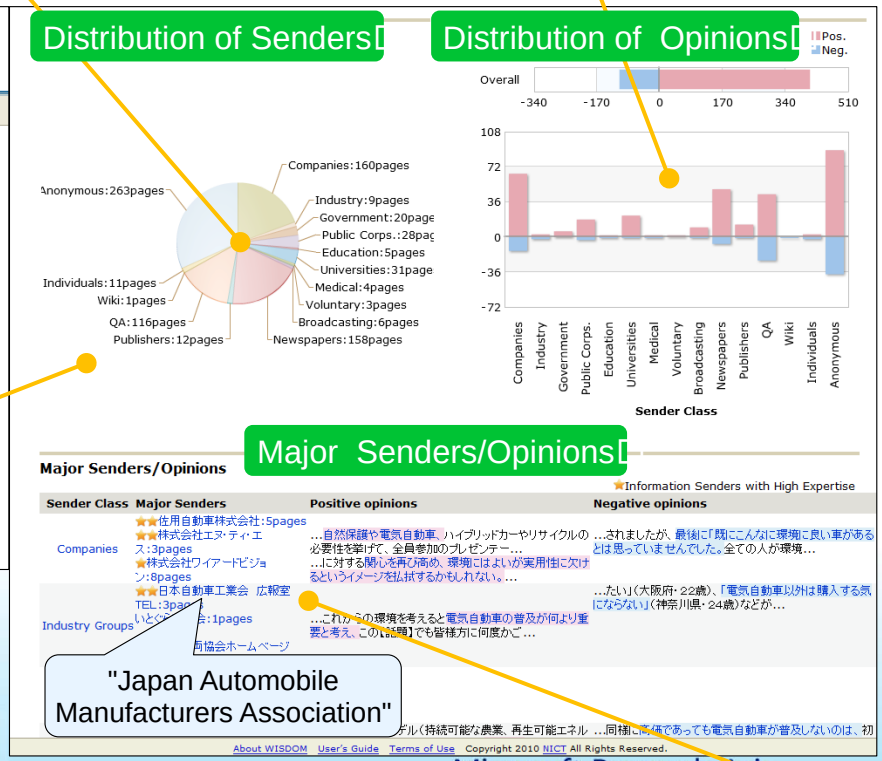
Input a topic to be analyzed.

Someone makes conflicting statements!

We can see major information sender classes!

The ratio of positive/negative opinions is different for each sender class!





We can grasp at a glance important issues and the distributions of information senders and opinions!

We can find experts on the topic!

Microsoft Research

Thank you!

Microsoft Research Asia
Faculty Summit 2012