

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



# Overview

Web Services

Search Engine  
TSUB

HPC or Cloud,  
Large Web Page Pool

Information  
on the Web

Web services of search  
and analysis based on  
deep NLP

Deep NLP-based  
indexing and search

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



Microsoft Research Asia  
**Faculty Summit 2012**



Web pages

Wikipedia

Deep NLP

Index

ねぎは

~~禰宜~~

葱 (syn.)

野菜 (hyp.)

食べ物 (hyp.)

風邪に

病気 (hyp.)

効果がある

効く (syn.)

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

Hypernyms  
Hyponyms  
500K pairs

Synonymous words/phrases  
50K pairs: 98% 30K pairs: 66%

Unknown words  
Hiragana Unk. detection: 34.5%→72.0%  
Unk. POS guess: 97.3%~98.4%

泳ぐ	ガ	生徒	彼女...
	デ	クロー	ル...
	ヲ	大海	海...
見る	ガ	子供	弟...

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



とろろ餅, 好酸球,  
灌がしい, ぱい焼, ...



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](#)

Article [Talk](#)

[Read](#) [Edit](#) [View history](#)

Search



# 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.<sup>[1]</sup> Lenovo purchased [IBM's](#) business and acquired the ThinkPad brand in 2005.

ThinkPads are popular with large businesses and government agencies. 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](#).<sup>[2]</sup>

## 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

Hypernym

## ThinkPad series

# ThinkPad



IBM ThinkPad R51

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



Know  
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 sandwich

Mary ate the sandwich

クロールで泳いでいる女の子を見た

crawl

swim

girl

saw



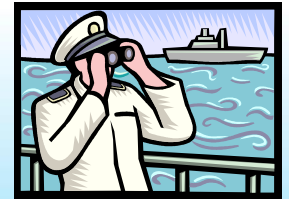
望遠鏡で泳いでいる女の子を見た

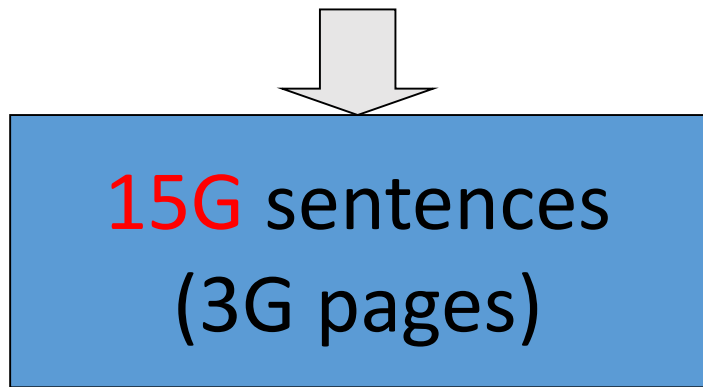
telescope

swim

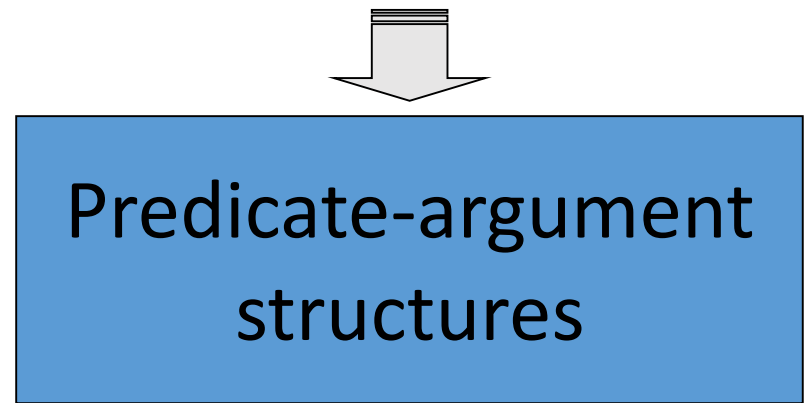
girl

saw

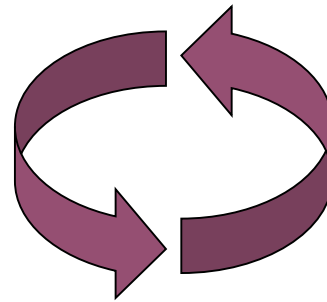




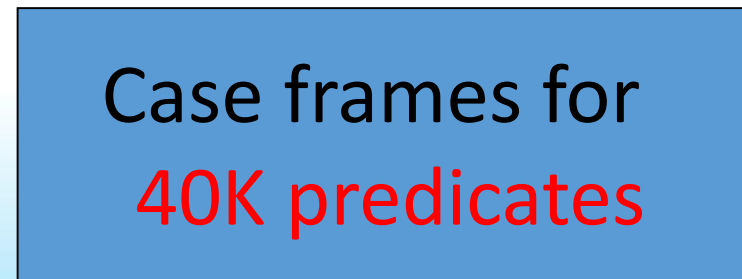
89.0% for all  
98.3% for 20.7% PAs



2weeks



2days



89.0% → 89.7%

Microsoft Research Asia  
Faculty Summit 2012



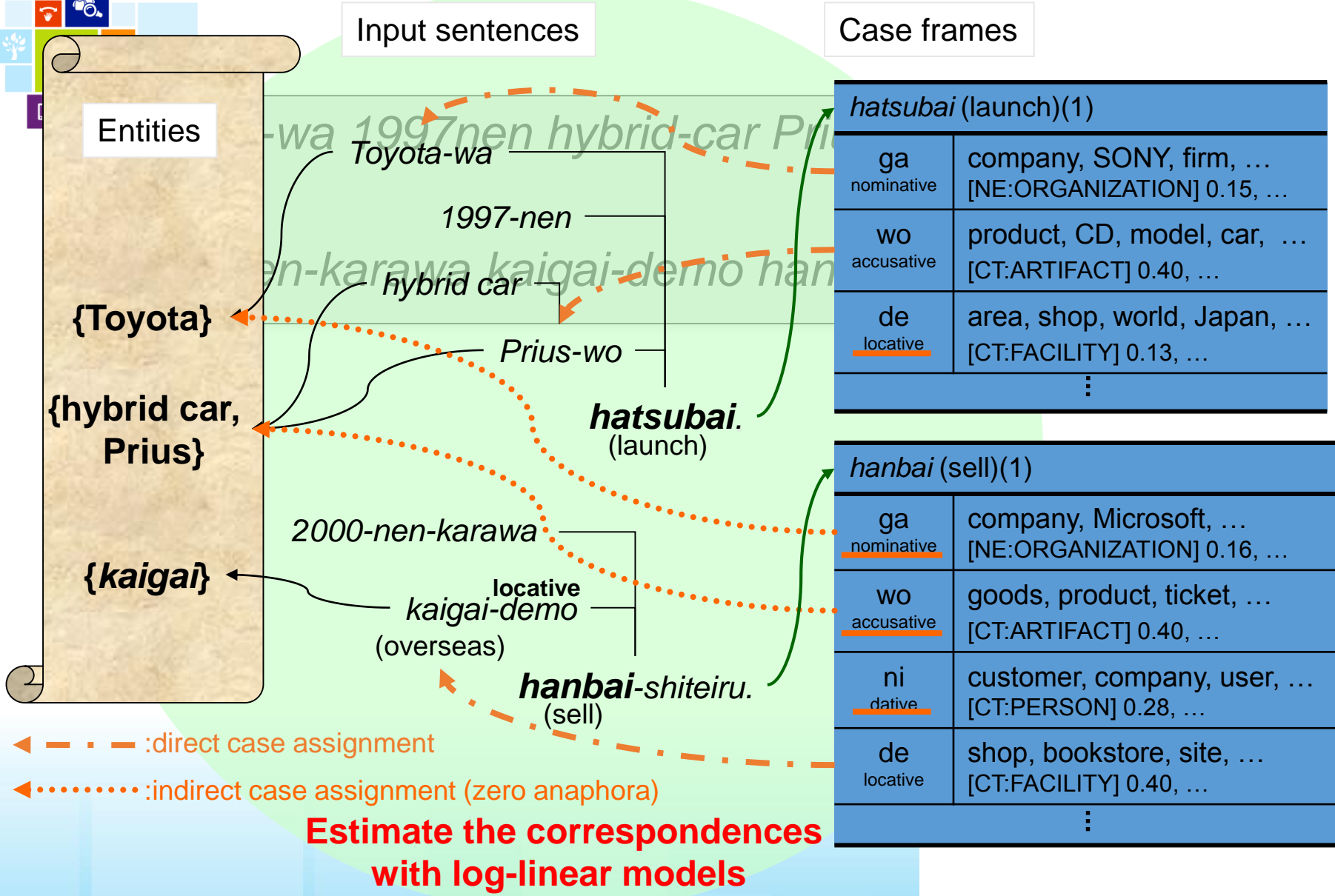


# Case frame examples

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

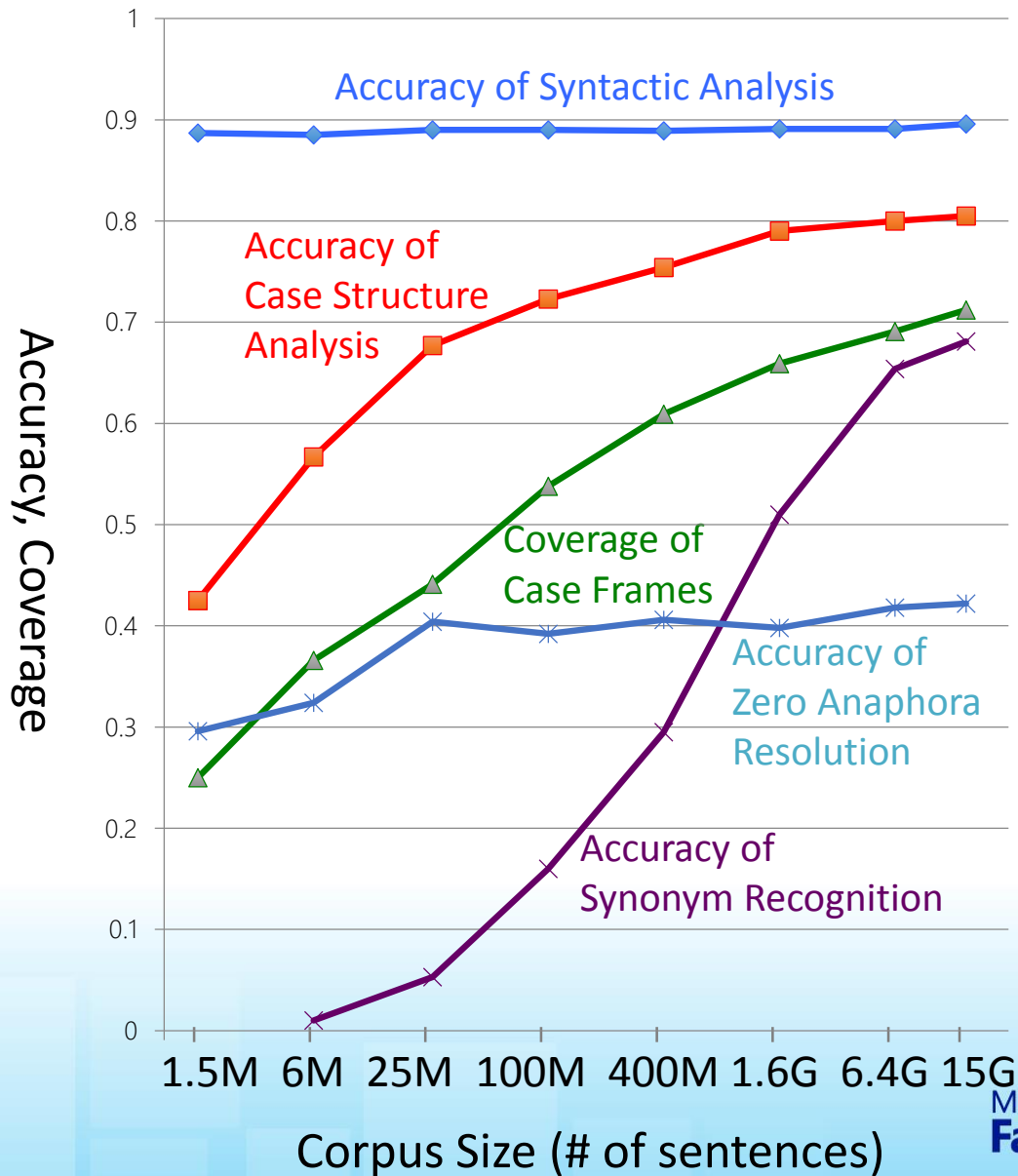


# Ellipsis (Zero Anaphora) Resolution



⊗ Toyota launched the hybrid car Prius in 1997.  $\phi_1$  started selling  $\phi_2$  overseas in 2000.

# Learning Curve





Web pages

Wikipedia

Deep NLP

Index

ねぎは

~~禰宜~~

葱 (syn.)

野菜 (hyp.)

食べ物 (hyp.)

風邪に

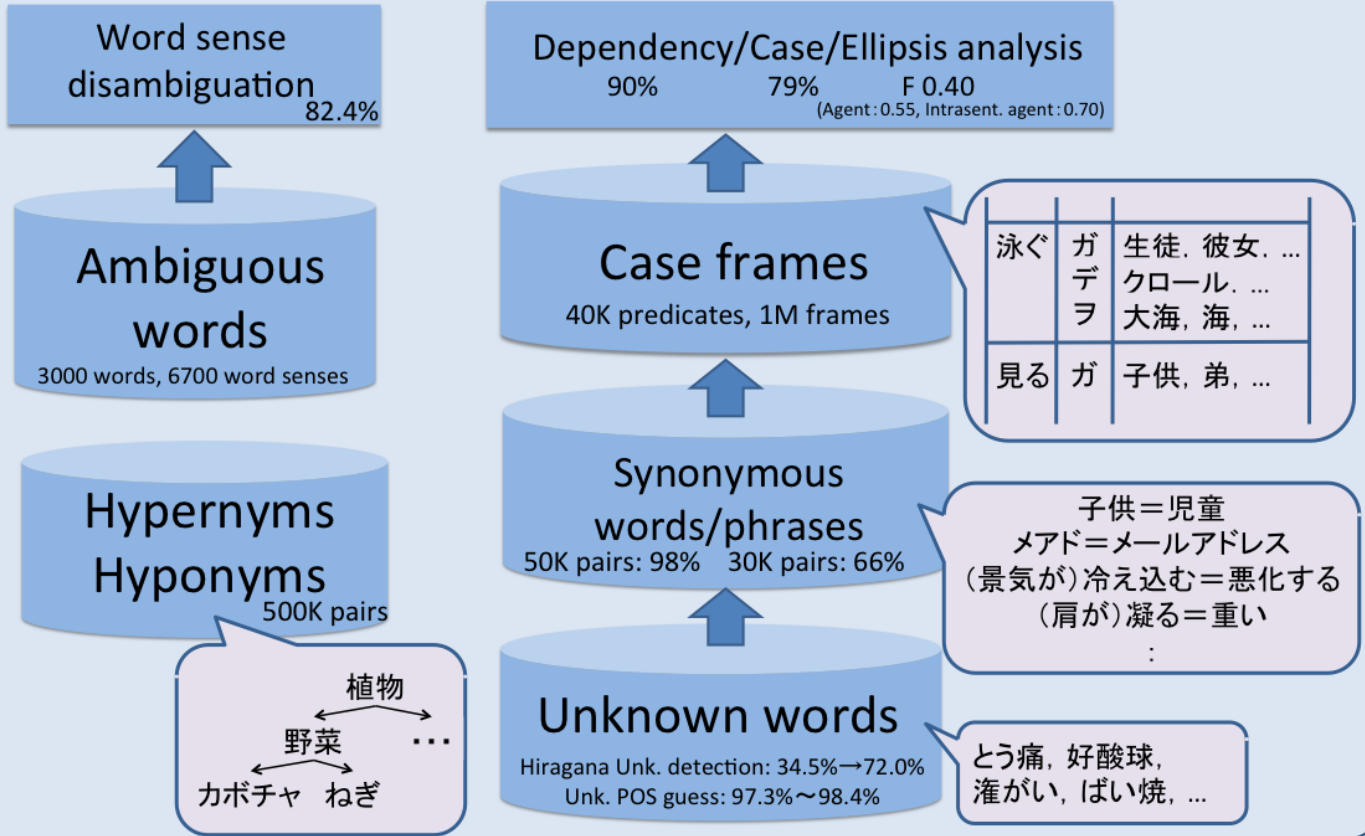
病気 (hyp.)

効果がある

効く (syn.)

Word sense disambiguation

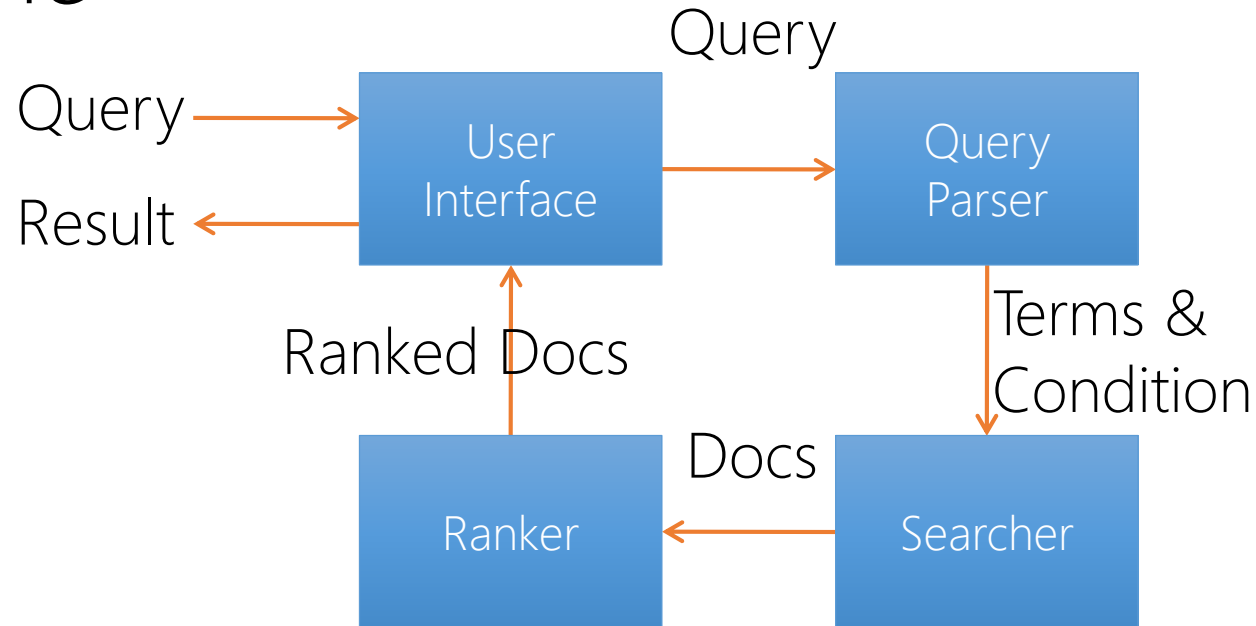
### Knowledge acquisition and deep NLP



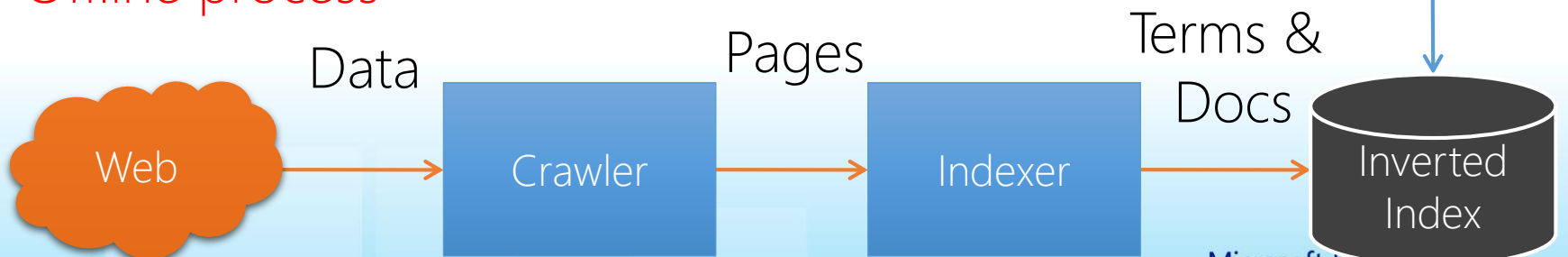


# Architecture of Search Engine

Online process



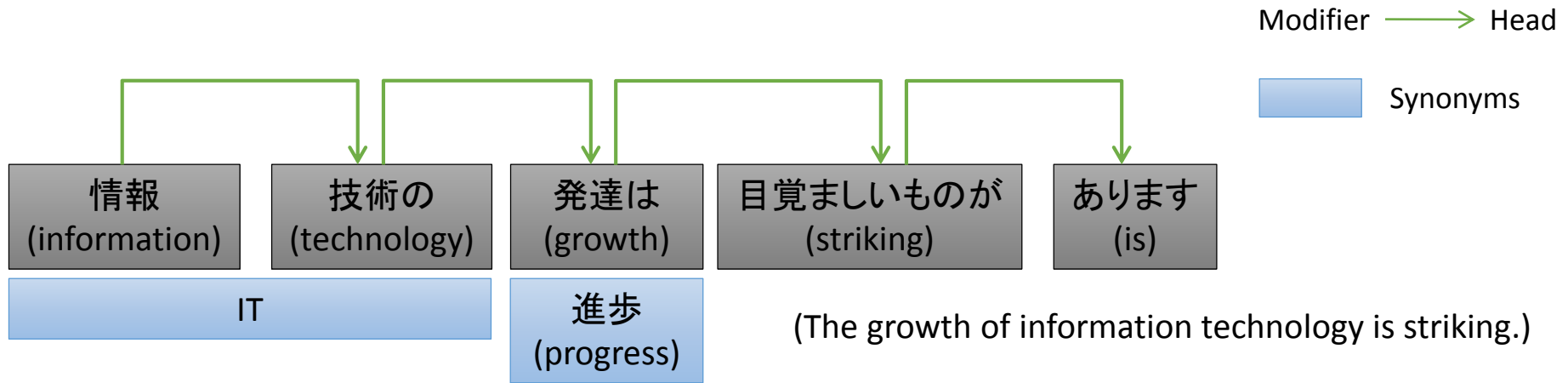
Offline process





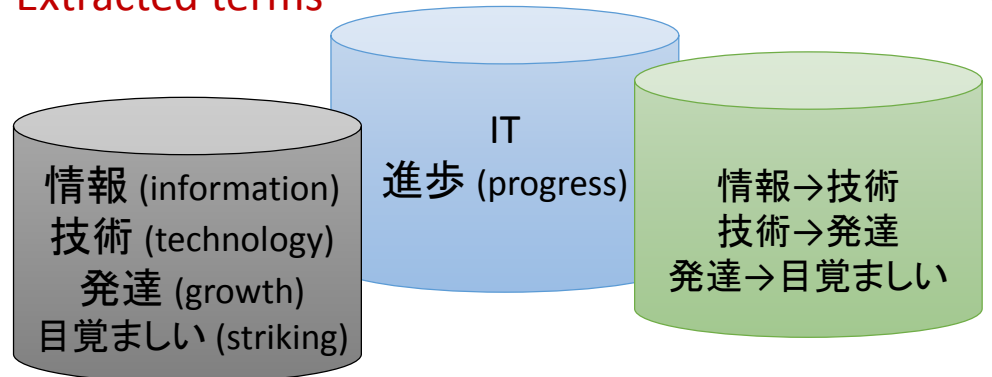
# Deep NLP-based Indexing

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



Indexing

Extracted terms





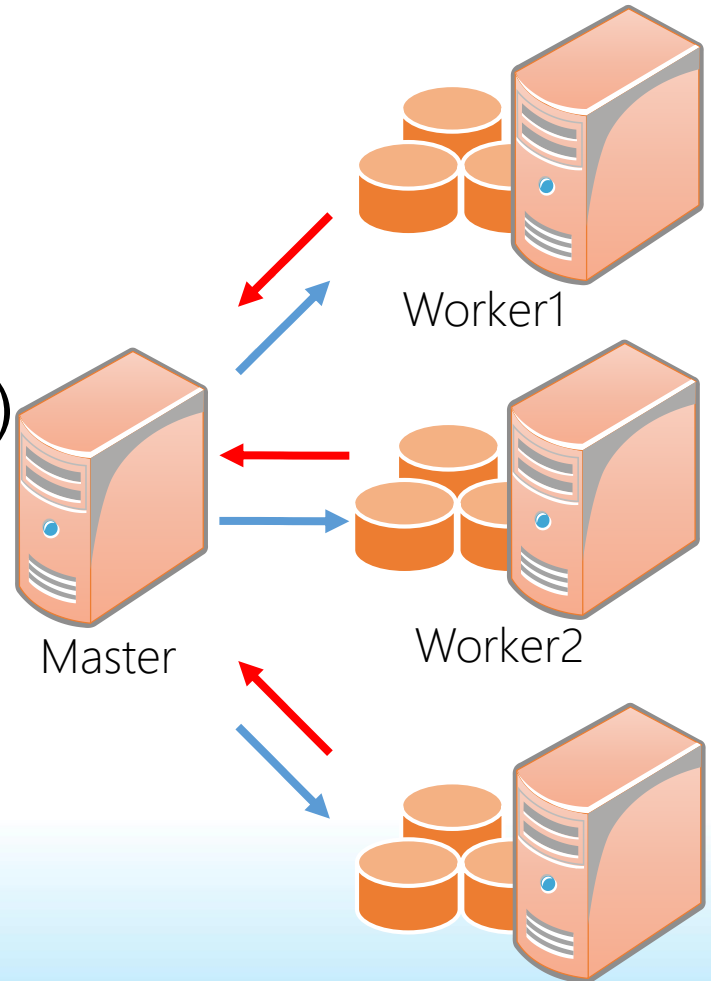
# Processing using HPC / Cloud

- 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

情報爆発プロジェクト 次世代サーチエンジン  
**TSUBAKI++**

Facebookが直面する問題  
(Problems that Facebook confronts)

1 [笑っている場合ですよ 利用者急増でフィッシング対策強化を強い...](#)

score = 98.8340

確かにリンク広告は苦痛だが、スパムは Facebook が直面する問題のなかで最も軽微に属するだろう。ユーザーが「ミスリードされてリンクをクリックした」セキュリティ上の問題になってくる。ユーザーの個人情報を守ることにかなりかかっている。

<http://waraimirai.blog109.fc2.com/blog-entry-3000.html>

... Spam may be the Facebook confronts.

2 [Facebook「ユーザーデータはユーザーのもの」ときっぱり...](#)

score = 97.8720

Googleが長年の間に收拾してきたわれわれの検索語か？ Facebookにアップされた情報はその何倍も個人 Facebookは攻撃するのにかっこうのターゲットであり結局は、ユーザーの怒りがまだ表面化していないことだ。

... This is the third affair that Facebook

3 [Facebookに「ガラスの顎」みたいな弱点あり](#)

score = 97.0320

【この記事はゲストのSteve Gillmorによる投稿です。ない重かな 難問に直面した。Googleの、見事なまで SNS 二人は、いずれユーザーたちの大きな怒りの津波に襲は、ユーザーたちの怒りがまだ表面化していないことだ。

... Facebook finally fa left. ...

4 [facebook - Toshiro Shimura on ...](#)

score = 95.1610

Facebookは大学生同士のつながりとして始まったため、ユーザー登録時に実名を使わないと信れず「つながりの中から外れていく」(ザッカーバーグ氏)という文化が当初からあったという。2Exp Facebookのユーザーベース全体としては、それほど懸念が生じたわけではなかったものの、

Google

Facebookが直面する問題 検索

約 61,000 件 (0.07 秒) 検索オプション

~~Facebook映画から学ぶ—中小ベンチャー企業が共通して直面する問~~

~~Facebook映画から学ぶ—中小ベンチャー企業が共通して直面する問題. by 小谷川 拳 (Kotanigawa Kenji) on Tuesday, January 25, 2011 at 7:40am. ひとたび社員に経営方針が浸透すれば、その企業は並々ならぬ力と柔軟性を発揮する ...~~

~~www.facebook.com/note.php?note\_id=158637594189130 - キャッシュ~~

~~日本の農業が抱える問題の一つが「分散錯圃... | Facebook ☆~~

~~Chikashi Horada この問題、金子勝さんも何度も話していました。高論こなればなるほど~~

~~www.facebook.com/mycastercom/posts/150289505028643 - キャッシュ~~

~~御社が抱える問題を、私たちが解決へお導き... | Facebook ☆~~

~~Sign up for Facebook helps you connect and share with the people in your life. ...~~

~~www.facebook.com/bbjpn/posts/198614263491530 - キャッシュ~~

~~facebook.com のその他の検索結果を表示する~~

~~リードコンサルティング社長 小谷川拳次のブログ Facebook映画から学~~

~~Facebook映画から学ぶ—中小ベンチャー企業が共通して直面する問題. 2011年01月~~

~~(株) 盛田昭夫 ひとたび社員に経営方針や理念が浸透すれば、その企業は並々ならぬ力~~

~~を発揮する —ソニー株式会社 代表 盛田昭夫 ...~~

~~le-tdcconsultingceomsg.blog89.fc2.com/blog-entry-62.html - キャッシュ~~

# Search Example: Use of P-A Structure



情報爆発プロジェクト 次世代サーチエンジン  
**TSUBAKI++**

## クラウドが解決すべき課題 (Problems that should be solved by Cloud)

検索する クリア

検索時間: 3.0 [秒]

1 [SMBにはSMB向け製品を——大企業の「お下がり」無用——](#)  
score = 258.8900 (w=0.000, d=0.000, n=0.000, aw=0.000, ad=0.000) [類似・関連ページを表示 \(42 件\)](#)  
クラウド 化とセキュリティは両立できる? クラウド サービス選択の新基準 中小企業にも使いやすいバックアップソフトウェアの要件 安全な仮想化環境の3条件とは——ブームの裏にセキュリティ崩壊の危機 儲からないネットショップの知られざる原因とは? 最も身近な危機“バンデミック”へのBCPを考える ホワイトペーパーBEST10 2009年12月21日更新  
<http://techtarget.itmedia.co.jp/it/news/0606/09/news07.html>

2 [これで日本は蘇る! 胎動する新事業とテクノロジー/Tech総研](#)  
score = 256.4820 (w=0.000, d=0.000, n=0.000, aw=0.000, ad=0.000)

There are **problems**. Extensibility and elasticity are the characteristics of **Cloud Computing**, but these have not reached the degree that companies require at the first priority. If **solved**, the barrier of cost is lowered and ...

すべて課題として以下のものがあつた。セキュリティの問題 複数種類と複数のバージョンのモデル ...  
<http://blog.livedoor.jp/bbreak/archives/51214068.html>

4 [プライベートクラウド構築・運用サービス「BusinessSt...](#)  
score = 165.9110 (w=0.000, d=0.000, n=0.000, aw=0.000, ad=0.000)  
プライベート クラウド 導入におけるポリシーにあった、物理リソースの配置設計、仮想化方式設計、予備方式設計、運用監視システム設計などのシステム設計を行います。システム構築サービス サーバ、ネットワーク、各々のリソース仮想化方式設計に基づいた構築をご提供します。運用監視システム設計・構築サー  
<http://www.hitachijoho.com/solution/outsourcing/bspp/>

5 [iPhoneやiPad、クラウド環境での印刷を実現! クラウド...](#)  
score = 165.6270 (w=0.000, d=0.000, n=0.000, aw=0.000, ad=0.000)  
クラウド・プリンティングセミナー事務局 本セミナーは定員に達したため受付を終了いたしました。今後ビジネス+ITセミナーをよろしくお願ひ申し上げます。セミナー名称 ... クラウド 環境で快適な印刷を実現 9月28日(火) 14:30~16:40 (14:00~ 受付開始) ...  
<http://www.sbbit.jp/eventinfo/11060>

絞り込みを解除

重要キーワード

[クラウド](#) [緊急報告](#) [特長](#) [コンピューティング II](#) [米国](#) [インフラ](#)  
[EC2](#) [Windows 7](#) [クラウドコンピューティング](#) [物理サーバ](#) [取り組み](#)

人・主体

[ユーザー](#) [ワランク](#) [パートナー](#)

組織・団体

[SaaS](#) [VMware](#) [オンデマンド型のERP](#) [Android](#) [外資のSAP](#)  
[Force.com](#) [IFRS](#) [セールスフォース](#) [Oracle](#) [ISV](#) [Lotus](#)  
[Notes](#)

その他

[ビジネス](#) [中小企業](#) [アプリケーション](#) [課題](#) [英国の6カ国](#) [国産のベンダー](#) [クラウド時代](#) [グループ会社の対応](#) [IFRS対応](#) [ERPのSaaS利用](#) [IFRSの会計処理](#)

### 解決 (solve)

ga	system, solution, company, expert, ...
wo	issue, <b>problem</b> , incident, trouble, conundrum, defect, ...
ni	early, actually, in turn, ...

# Search Example: Hypernym/Hyponym Relations

風邪の予防に効果的な野菜  
(Effective vegetable to the prevention of the cold)

Hyponym of vegetable

Chinese cabbage: One of the brightly colored vegetables that plenty contains vitamin A... as an effective vegetable that contribute to prevent the cold.

Hyponym of vegetable

Welsh onion is perfect for preventing the cold because it has effects for warming your body, facilitating the circulation, enhancing appetite and improving immunity.

Hyponym of vegetable

The recovery effect can be obtained when eating pumpkins that include vitamin A. Vitamin A works to prevent skin roughness and the cold.

The screenshot shows a search engine interface with the following elements:

- Search engine: 情報爆発プロジェクト 検索エンジン基盤 TSUBAKI
- Search results for "風邪の予防に効果的な野菜" (Effective vegetable to the prevention of the cold).
- Search filters: Title, Keyword, Header, Footer, Link, etc.
- Search time: 5.3 [秒]
- Results list:
  - 1 チンゲンサイ (Chinese cabbage) with a callout: "Chinese cabbage: One of the brightly colored vegetables that plenty contains vitamin A... as an effective vegetable that contribute to prevent the cold."
  - 2 Yahoo! ヘルスケア-冬を元気に過ごそう (Welsh onion) with a callout: "Welsh onion is perfect for preventing the cold because it has effects for warming your body, facilitating the circulation, enhancing appetite and improving immunity."
  - 3 健康教室 (Pumpkin) with a callout: "The recovery effect can be obtained when eating pumpkins that include vitamin A. Vitamin A works to prevent skin roughness and the cold."
  - 4 全国郷土料理レシピ
  - 5 none



# 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	<b>0.173</b>	<b>0.442</b>	<b>0.379</b>	<b>0.237</b>

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! □

**Major/Contradictory Statements**

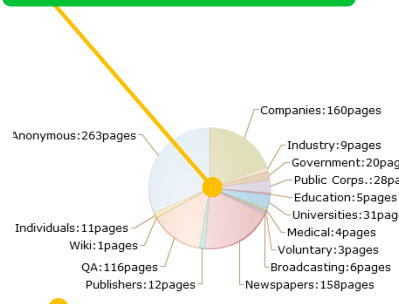
**Major Keywords**

**"CO2", "fuel consumption", "exhaust gas", ...**

**"good for the environment"**

**"not good for the environment"**

### Distribution of Senders



### Distribution of Opinions



### Major Senders/Opinions

Sender Class	Major Senders	Positive opinions	Negative opinions
Companies	★株式会社エヌ・ティ・エス:3pages	★自然保護や電気自動車、ハイブリッドカーやリサイクルの...	★「最後に「類」にこんなに環境に良い車がある必要は...」
Industry Groups	★日本自動車工業会 広報室:1pages	★「これからの環境を考えると電気自動車の普及が何より重要と...」	★「(大阪府・22歳)、「電気自動車は購入する気にならない」(神奈川県・24歳)などが...

"Japan Automobile Manufacturers Association"

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

We can find experts on the topic! □

# Thank you!