

Deep Natural Language Processing for Improving a Search Engine Infrastructure using Windows Azure

Daisuke Kawahara and Sadao Kurohashi

Graduate School of Informatics,
Kyoto University

Overview

Web Services

Web services of search and analysis based on deep NLP

Search Engine Index

TSUBAKI

Information
on the Web

Deep NLP-based
indexing and search

HPC or Cloud,
Large Web Page Pool



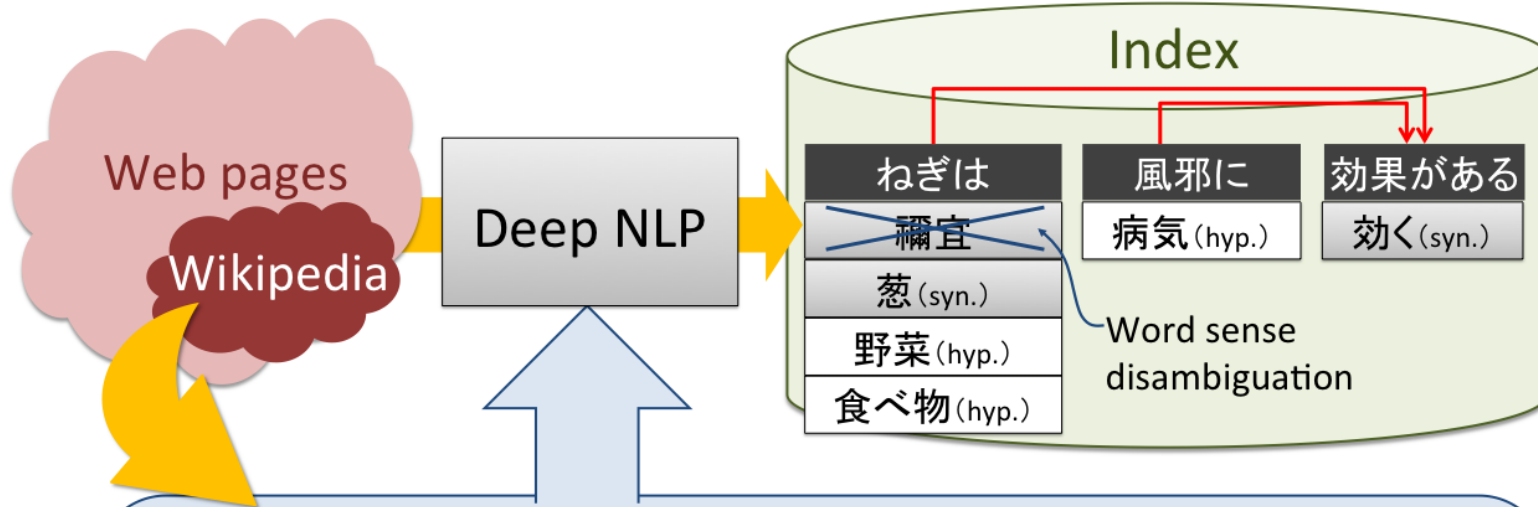
Windows Azure

- Linux clusters/grids
 - 1,000~2,000 CPUs
- Windows Azure
 - ~**10,000**~ CPUs

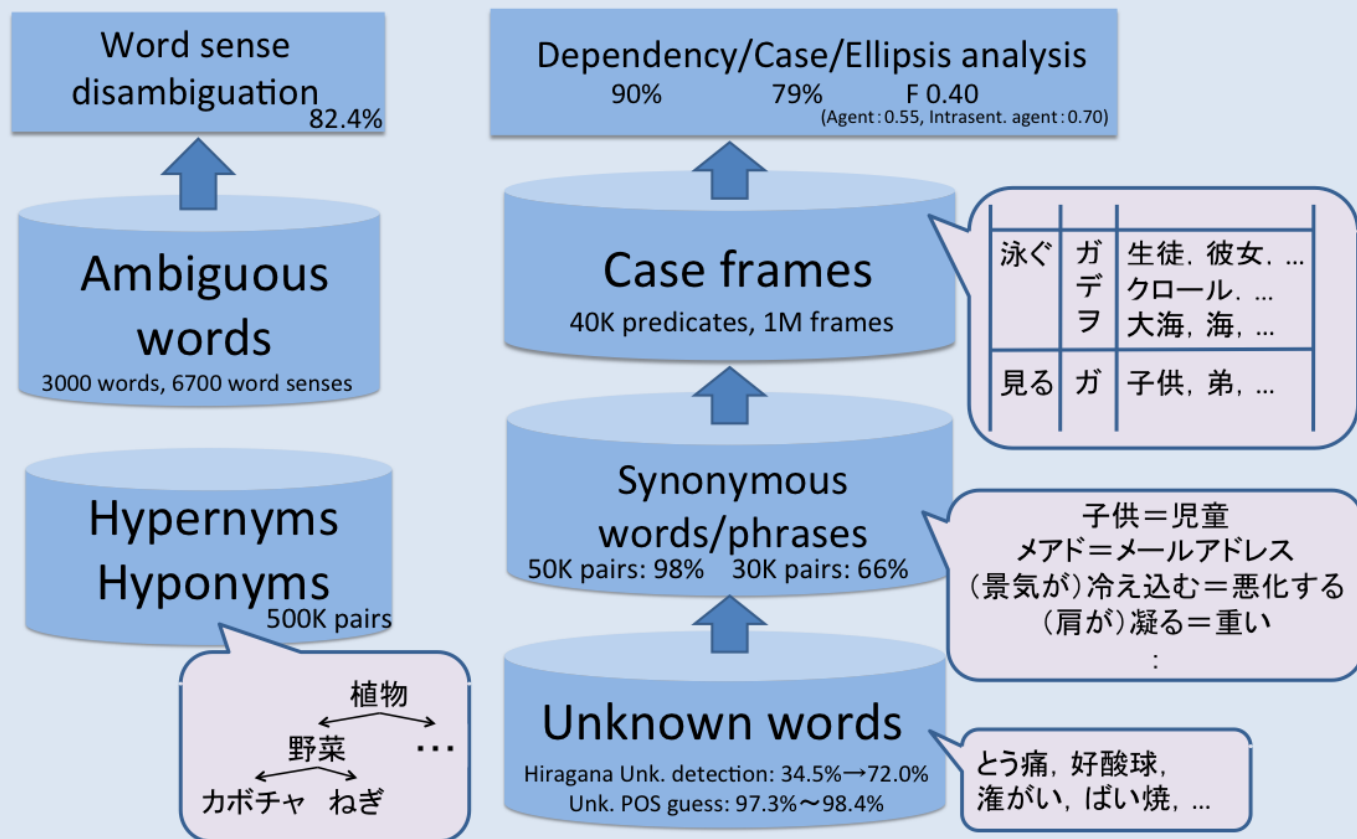
Search Engine Infrastructure

TSUBAKI

based on Deep NLP



Knowledge acquisition and deep NLP



Know Case frame

泳ぐ swim

{人 person, 子 child,...}が
{クロール crawl, 平泳ぎ,...}で
{海 sea, 大海,...}を

Mary ate the sa

見る see

Mary ate the sa

{人 person, 者,...}が
{望遠鏡 telescope, 双眼鏡,...}で
{姿 figure, 人 person,...}を

クロールで泳いでいる女の子を見た

crawl

swim

girl

saw



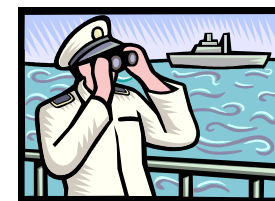
望遠鏡で泳いでいる女の子を**見た**

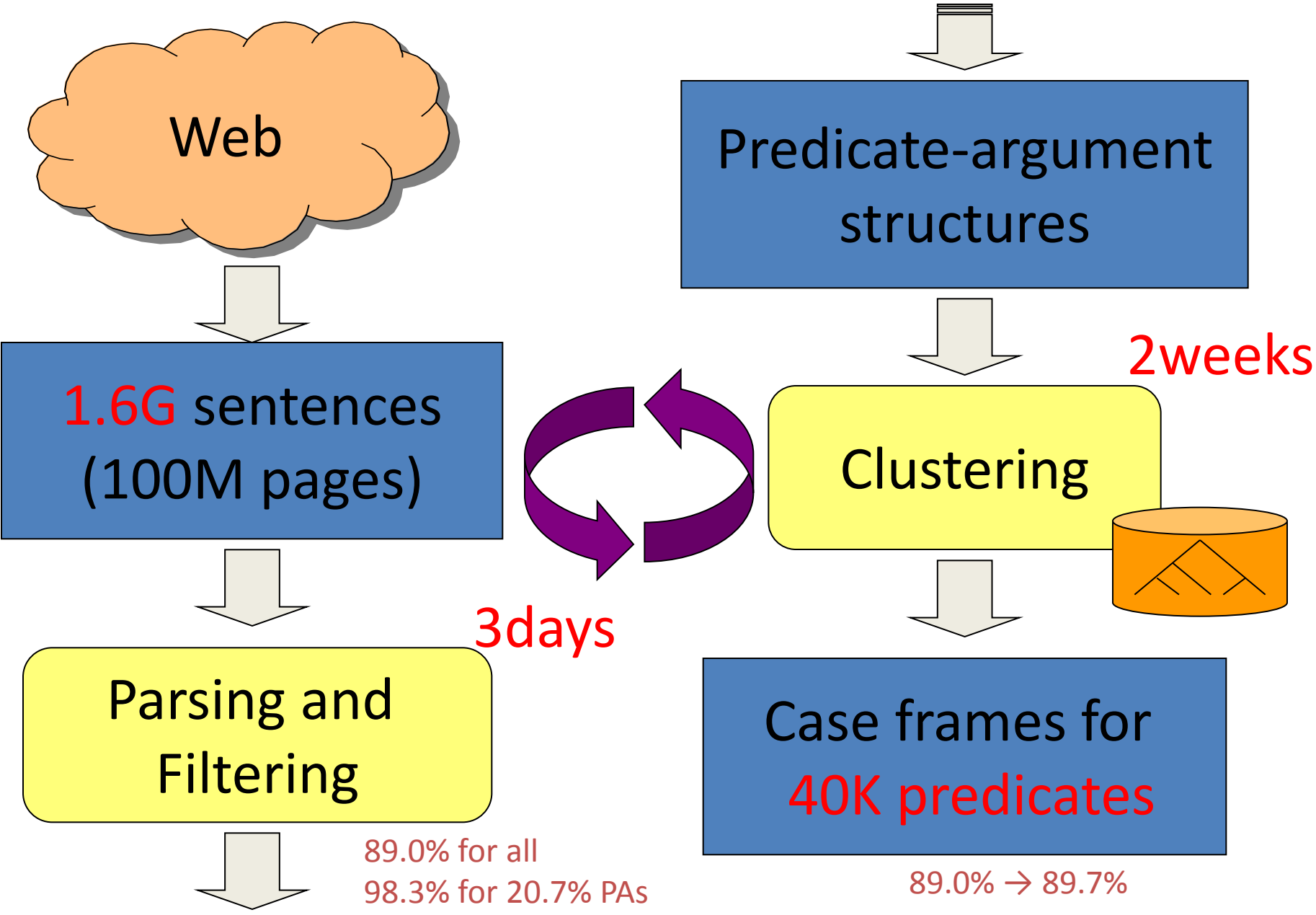
telescope

swim

girl

saw

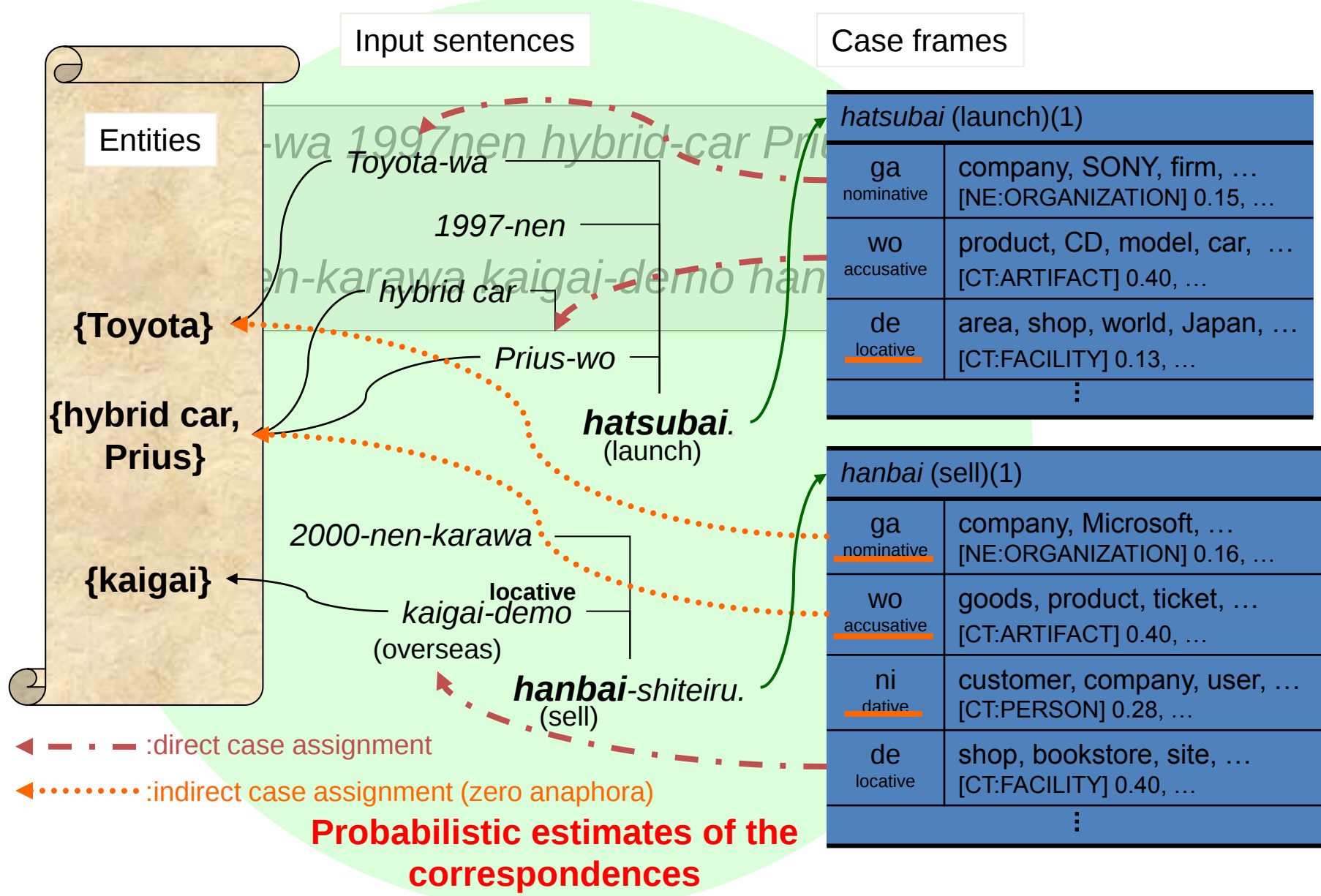




Case frame examples

	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, ...
...		

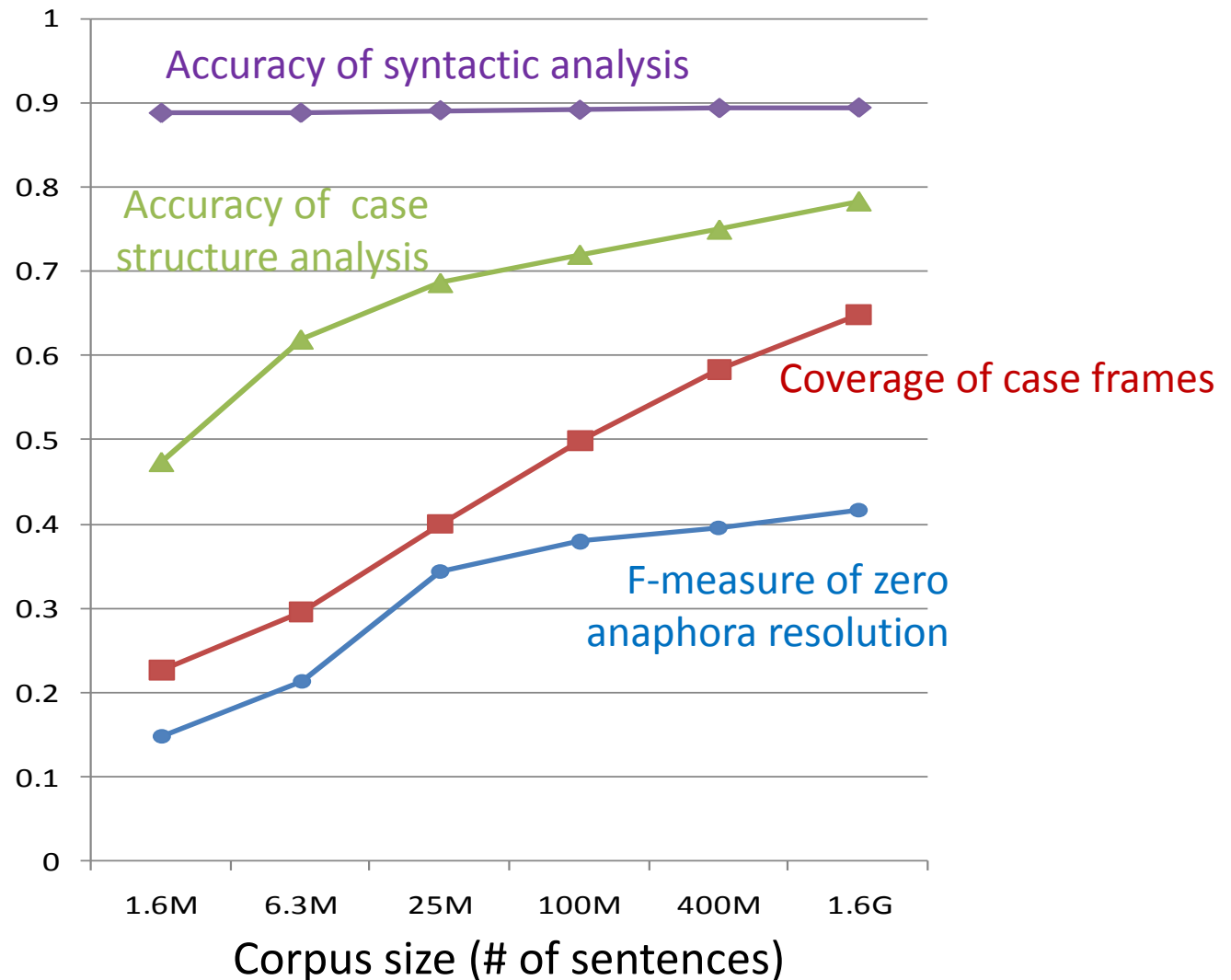
Ellipsis (Zero Anaphora) Resolution [Sasano et al., COLING08]



Toyota launched the hybrid car Prius in 1997. ϕ_1 started selling ϕ_2 overseas in 2000.

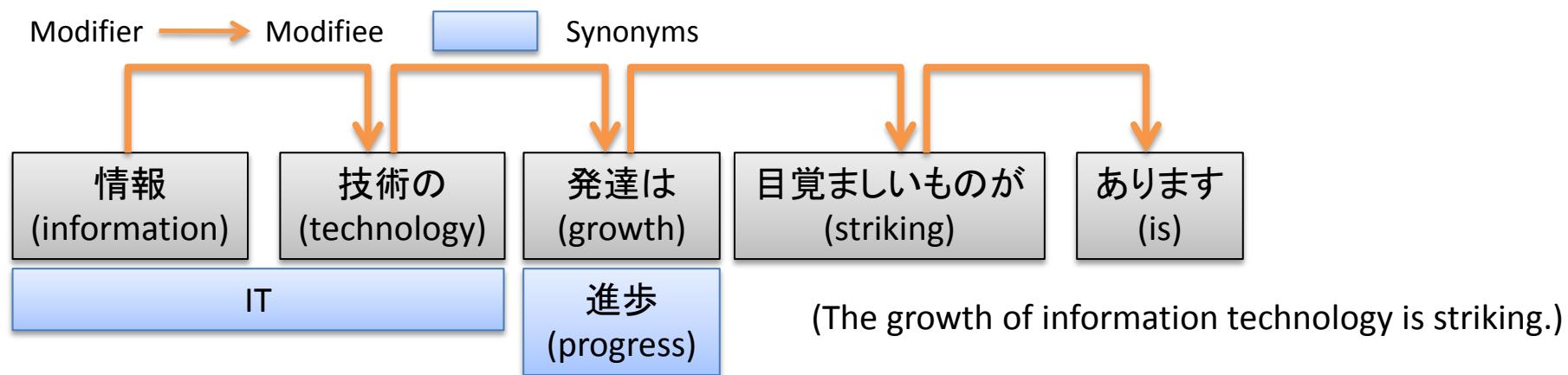
Effect of Corpus Size

[Sasano et al., NAACL-HLT09]

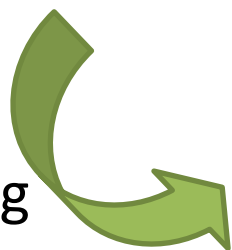


Indexing

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



Indexing

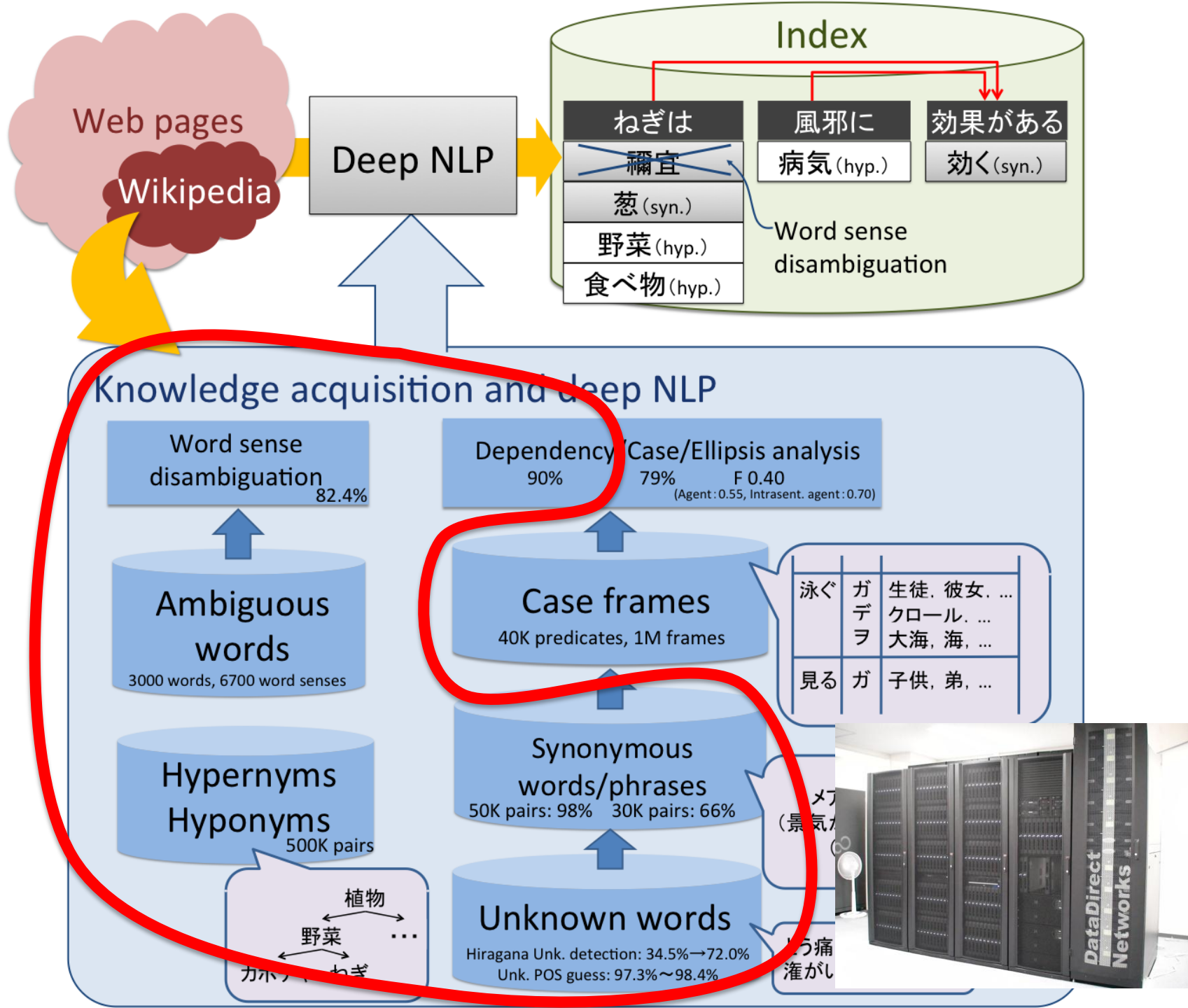


Extracted terms

情報 (information)
技術 (technology)
発達 (growth)
目覚ましい (striking)

IT
進歩 (progress)

情報→技術
技術→発達
発達→目覚ましい



Search Example: Use of Dependency Relations

情報爆発プロジェクト 次世代サーチエンジン

TSUBAKI++

Facebookが直面する問題で検索

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 = 18.8340
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
ヤセキュリティ上の 問題 になってくる。ユーザーの個人情報
を負わせることになりかねない。
<http://waraimirai.blog109.fc2.com/blog-entry-...>

2 [Facebook「ユーザーデータはユーザーのもの」ときっぱり...](#)
score = 17.8720
Google が長年の間に収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
<http://jp.techcrunch.com/archives/20090221...>
but-see-you-at-the-n...

3 [Facebookに「ガラスの顎」みたいな弱点あり](#)
score = 17.0320
【この記事はゲストのSteve Gillmorによる投稿です。
ない重大な 難問 に直面した。Googleの、見事なまで
SNS 世人は、いずれユーザたちの大きな怒りの津波に
は、ユーザたちの怒りがまだ表面化していないことだ。
<http://jp.techcrunch.com/archives/20080517...>

4 [facebook - Toshiro Shimura on ...](#)
score = 15.1610
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(ザッカーバーク氏)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なわけではなかったものの、

... Spam may be the
Facebook confronts.

... This is the third jaw
affairs that Facebook

... Facebook finally fa
left. ...

Facebookが直面する問題

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
案が浸透すれば、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](#) ☆ [Q](#)
Chikashi Horada この問題、金子勝さんも何度も話していました。高論になればなるほど
www.facebook.com/ycastercom/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](#) ☆ [Q](#)
Sign in 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](#)
☆ [Q](#)

[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
(木) 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 ...
leadconsultingceomsg.blog89.fc2.com/blog-entry-62.html - キャッシュ

Search Example: Hypernym/Hyponym Relations

情報爆発プロジェクト 検索エンジン基盤 TSUBAKI - Mozilla Firefox

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.

情報爆発プロジェクト 検索エンジン基盤 TSUBAKI

検索時間: 5.3 [秒]

1 チンゲンサイ
score = 89.7604 (w=36.329, d=16.852, n=36.579, av=0.000, ad=0.000)
■野菜 事典 ■野菜の特徴、効能...
れる事が多いため、1回あたりの摂取量が多く、風邪を
買う時は、みずみずしく、葉と茎にはりあるものを
http://www6.airnet.ne.jp/grama/cook/to

2 Yahoo!ヘルスケア-冬を元気に過ごそう
score = 76.6077 (w=31.806, d=11.926, n=32.876, aw=0.000, ad=0.000)
寒い季節には体の芯(しん)から温まりたいもの。風邪 ぎみのときにオススメのあったかレシピをご紹介します。サフランのリゾット245.0kcal(1人分)60分 風邪に 効果
的なサフランを使ったヘルシーメニュー。野菜のポタージュ100.6kcal(1人分)15分体力回復に効く乳製品と野菜を食べやすい形に。抵抗力の快復にはビタミン補給が 効
果 的。乳製品とのコンビで栄養価も味もアップ。梅にゆうめん216.0kcal(1人分)15分 ときにもピッタリなのがこの一品。梅干しの酸味でさっぱりといいただけま
す。すき煮うどん301.1kcal(1人分) ... スタミナ源を体が温まる ... ばく質が豊富。熱い汁と ねぎ がからだを温めます。チンゲンツァ
イのミルクおじや307.8kcal(1人分) ... 高栄養の牛乳と野菜
http://health.yahoo.co.jp/column/winter/recipe_01.htm

3 健康教室
score = 75.3805 (w=27.093, d=13.843, n=34.444, av=0.000, ad=0.000)
ここでは、保健婦かあばんが健康に関する話題を
ントです 風邪を 予防しましょうお手軽健康食レシピ
お手軽健康食レシピ集(その4)簡単で健康的アロ
http://www.takenet.or.jp/~hayakawa/ky

4 全国郷土料理レシピ
score = 73.5633 (w=27.689, d=12.340, n=33.534, aw=0.000, ad=0.000)
所変われば料理も変わる、日本全国都道府県の外
かせない ねぎ 料理だったそうです。『めた』とは酢
にはもってこいの 野菜 です。旬のこの時期、たっ
水分を切り、2)に混ぜ合わせ、器に盛って出来上
http://www.s-recipe.com/local/vol70.htm

5 none
score = 70.5897 (w=28.429, d=13.225, n=28.936, aw=0.000, ad=0.000)
その目に効くビタミンAを含んだ カボチャ を食べて、回復 効果 を得たのでしょう。またビタミンAには、同時に肌荒れを防ぎ、風邪の 予防 にもなるという役割もあります。□TOP

Web Services on TSUBAKI

Information Analysis with WISDOM

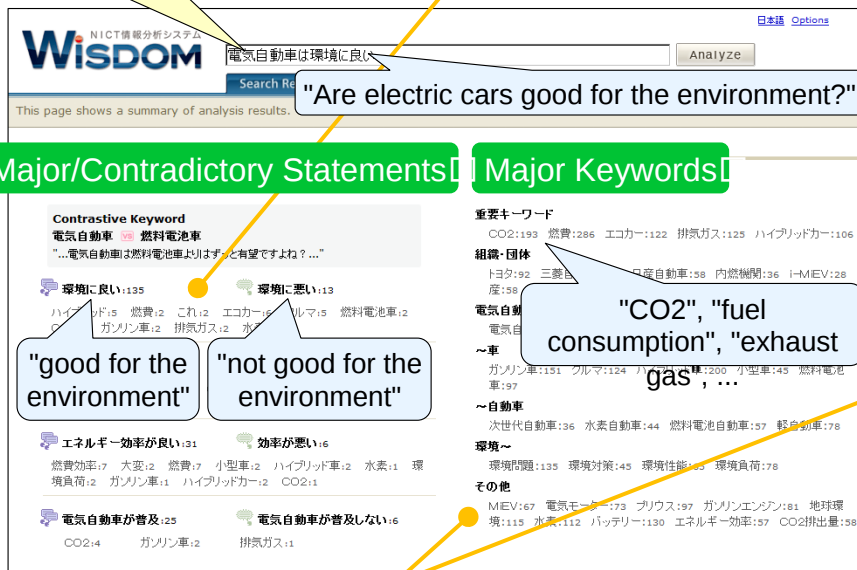
Analyzing web pages based on various criteria

Input a topic to be analyzed.

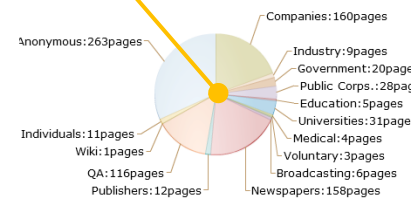
Someone makes conflicting statements!

We can see major information sender classes!

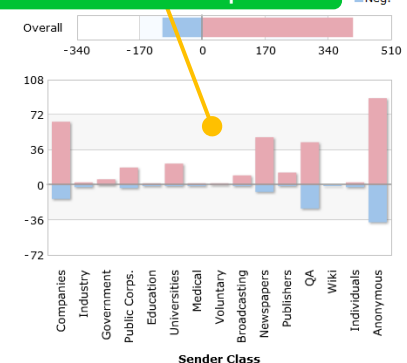
The ratio of positive/negative opinions is different for each sender class!



Distribution of Senders



Distribution of Opinions



Major Senders/Opinions

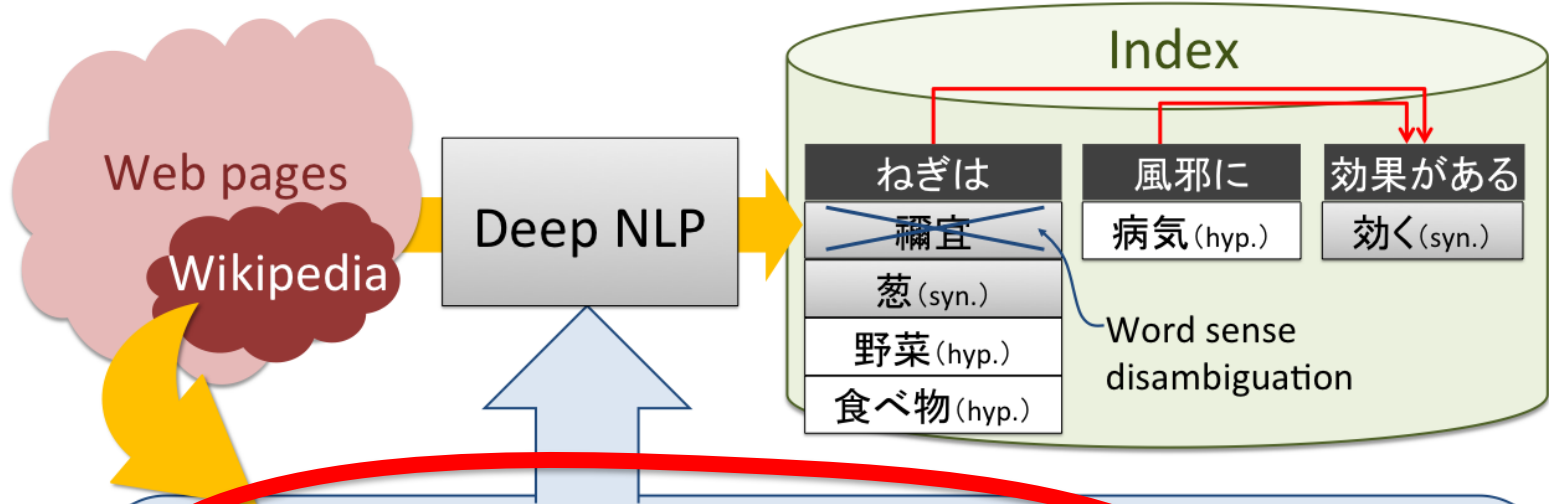
Major Senders/Opinions	
Sender Class	Major Senders
	Positive opinions
Companies	★住用自動車株式会社: 3pages
	★株式会社エス・ティ・エス: 3pages
Industry Groups	★株式会社ワイアードビジョン: 8pages
	★日本自動車工業会 広報室: 1pages
	Negative opinions
	★自然保護や電気自動車、ハイブリッドカーやリサイクルの... 最後は「環境にいい車がある」と思っていました。全ての人が環境... 必要を挙げて、全員参加のプレゼン... に対する期待の裏面、環境はよいが実用性に欠けるというイメージを払拭するかもしれない...
	★「たいい」(大阪府・22歳)、「電気自動車以外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!

TSUBAKI based on Deeper NLP
Enabled by Windows Azure



Knowledge acquisition and deep NLP

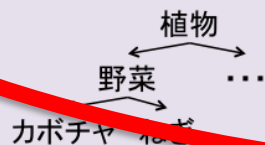
Word sense
disambiguation
82.4%

Ambiguous
words

3000 words, 6700 word senses

Hypernyms
Hyponyms

500K pairs



Dependency/Case/Ellipsis analysis
90% 79% F 0.40
(Agent: 0.55, Intracent. agent: 0.70)

Case frames

40K predicates, 1M frames

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
(肩が)

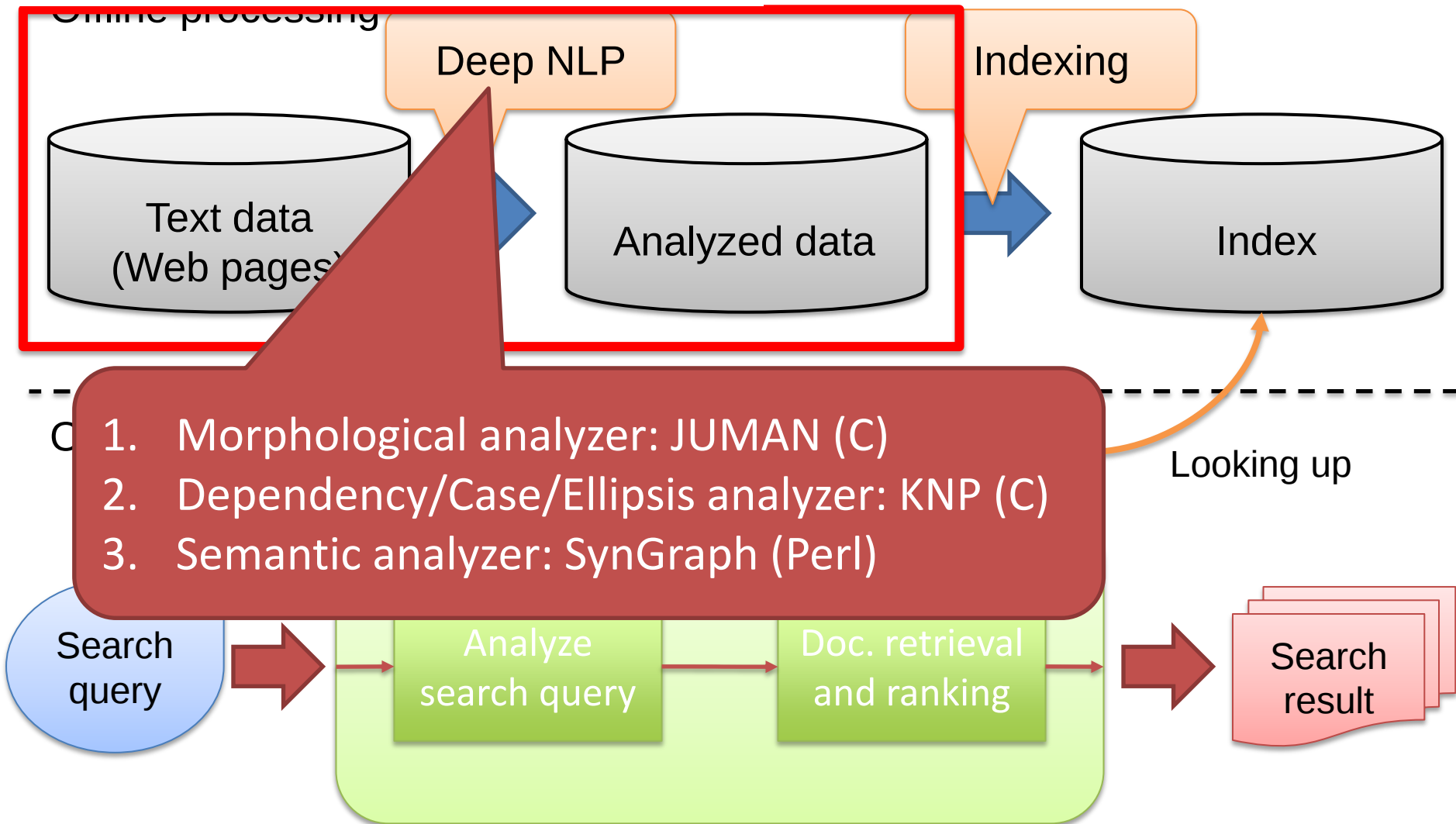
とうとう、好
づかしい、は



Windows Azure

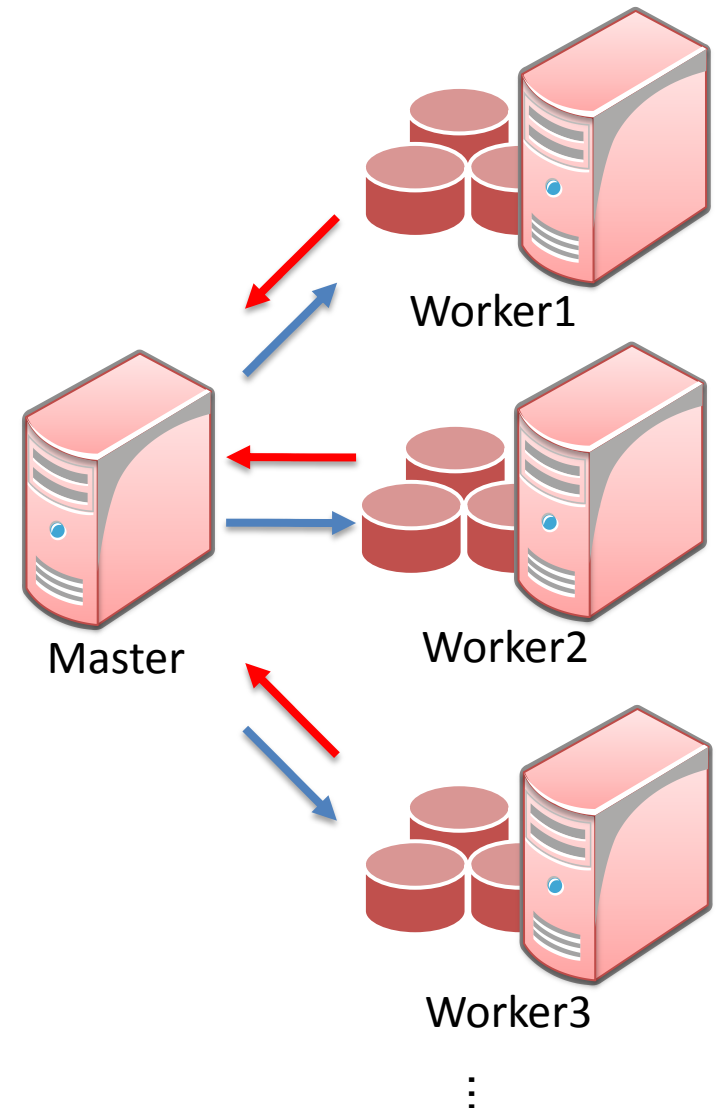
Flow of TSUBAKI

Execute deep NLP on Azure

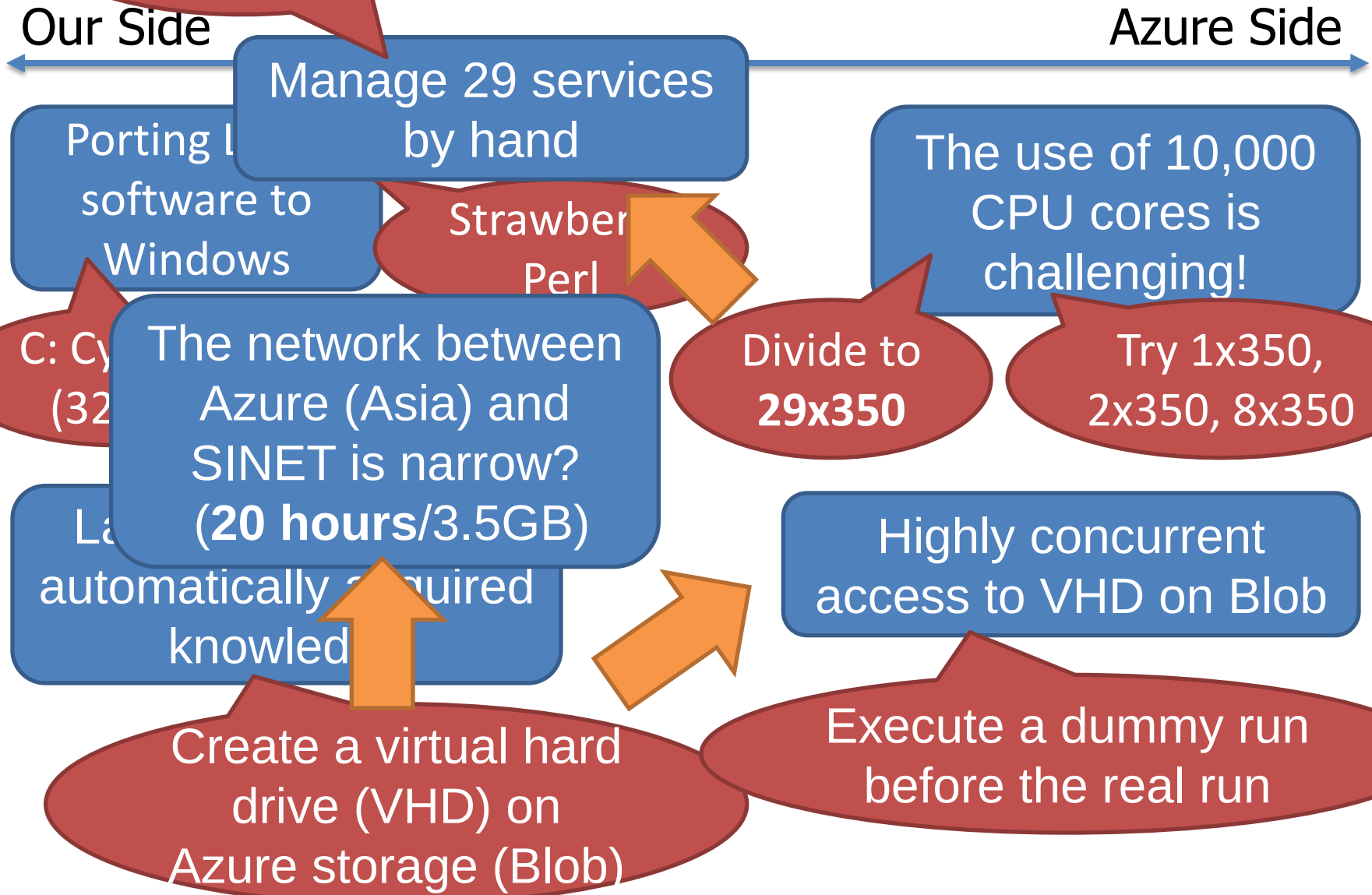


Our Model of Deployment to Azure

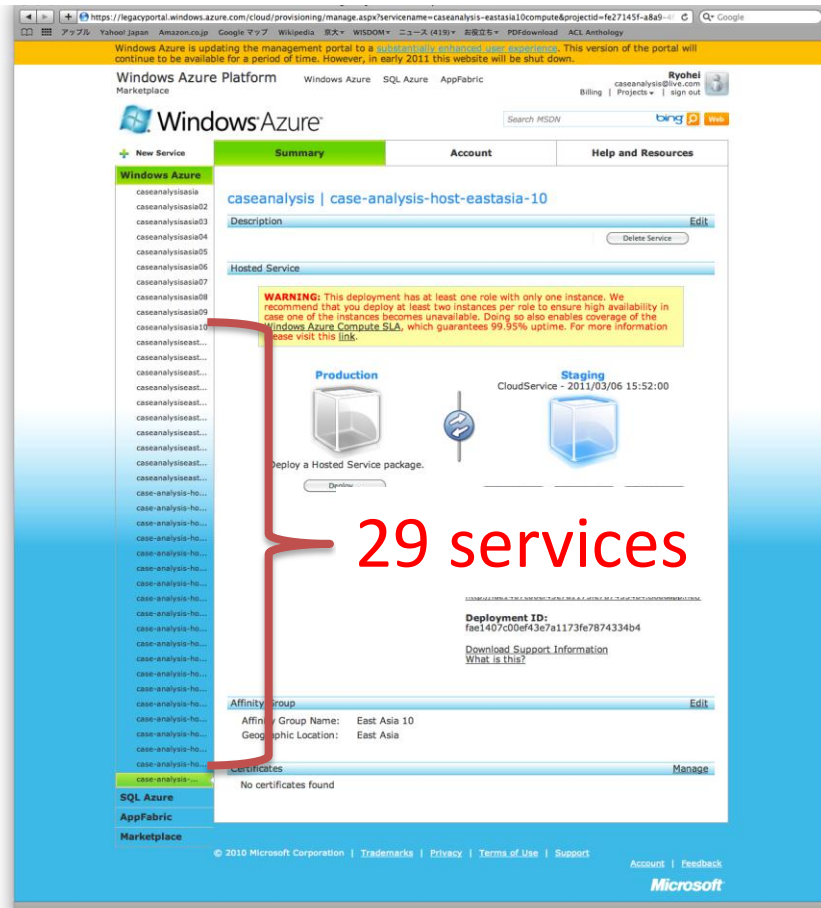
- Task is completely independent
 - Each task processes a small set of Web pages (100 pages)
- Master/Worker model
 - Master = Web Role
 - Worker = Worker Role



Problems and Solutions



Finally, we succeeded in using 10,000 CPU cores!



- TODO: real run
 - Upload all the input data (120M pages; 260GB)
 - Execute the analysis
 - Download the output data (3TB)

Search Example: Use of Case/Ellipsis Analysis

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 で検索された文書

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

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——プ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“パンデミック”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 ...

3 [クラウド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——プ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“パンデミック”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://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—
<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, problem , incident, trouble, conundrum, defect, ...
ni	early, actually, in turn, ...

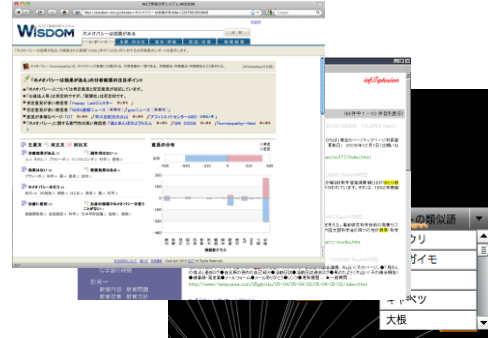
:

Conclusion

Web Services

Search Engine Infra
T S U B A K I

HPC or Cloud,
Large Web Page Pool



Web services of
search and analysis
based on deep NLP



Deep NLP-based
indexing and search

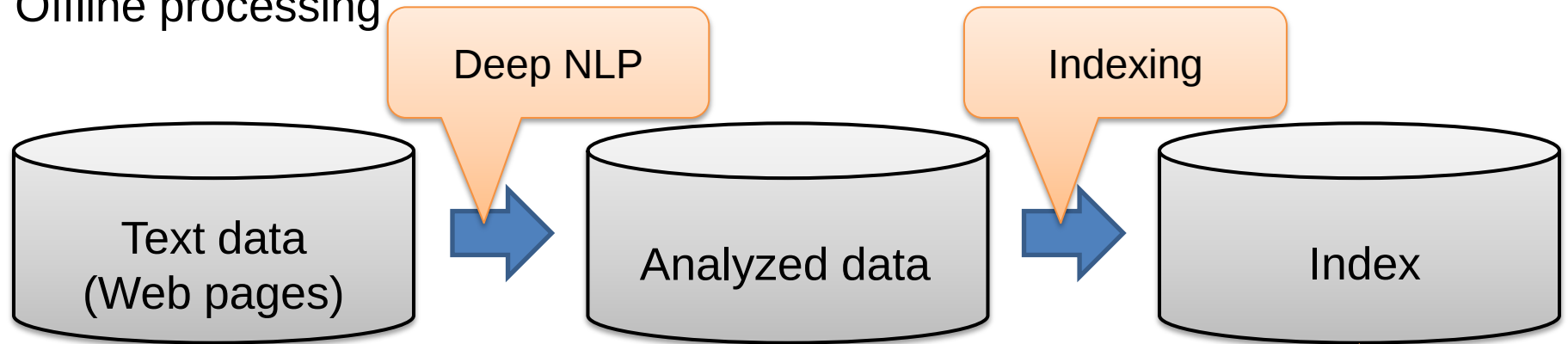


Windows Azure

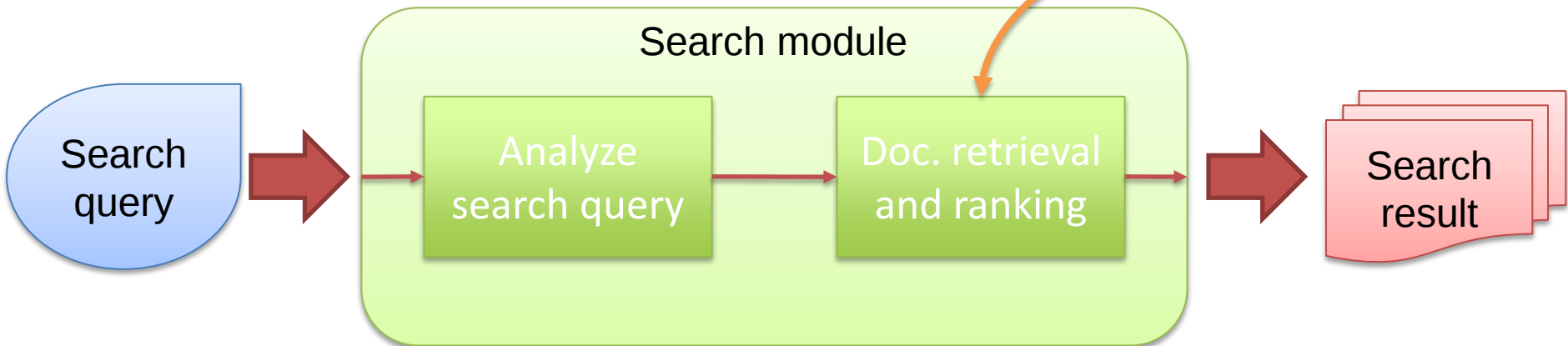
- Windows Azure
 - ~ **10,000** ~ CPUs

Future Work: **Execute the Whole Process on Azure!**

Offline processing



Online processing



Many thanks to

- Dr. Dennis Gannon (MSR)
- Ms. Hitomi Iida (MS Support)
- Prof. Masaru Kitsuregawa (Univ. of Tokyo)
- Prof. Jun Adachi (NII)



Windows® Azure™