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

Daisuke Kawahara and Sadao Kurohashi

Graduate School of Informatics, Kyoto University

Cloud Futures Workshop 2011 (June 3rd, 2011)
Overview

Web services of search and analysis based on deep NLP

Deep NLP-based indexing and search

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

Information on the Web

Web Services

Search Engine Infra

T S U B A K I

HPC or Cloud, Large Web Page Pool
Search Engine Infrastructure

TSUBAKI

based on Deep NLP
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
  - 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%

- **Index**:
  - ねぎは (hyp.)
  - 禁宜
  - 葱 (syn.)
  - 野菜 (hyp.)
  - 食べ物 (hyp.)
  - 風邪に (hyp.)
  - 効果がある (syn.)

- **Web pages**: Wikipedia

**Word sense disambiguation**
Knowledge Acquisition and Deep NLP

Mary ate the salad with a fork.

Mary ate the salad with mushrooms.

クロールで泳いでいる女の子を見た

望遠鏡で泳いでいる女の子を見た
Case frames for 40K predicates

1.6G sentences (100M pages)

Web

Predicate-argument structures

Clustering

Parsing and Filtering

3 days

89.0% for all
98.3% for 20.7% PAs

89.0% → 89.7%

2 weeks

[Kawahara and Kurohashi, LREC06, HLT-NAACL06]
## Case frame examples

<table>
<thead>
<tr>
<th>yaku (1) (bake)</th>
<th>CS</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>ga</strong></td>
<td>I:18, person:15, craftsman:10, ...</td>
</tr>
<tr>
<td></td>
<td><strong>wo</strong></td>
<td>bread:2484, meat:1521, cake:1283, ...</td>
</tr>
<tr>
<td></td>
<td><strong>de</strong></td>
<td>oven:1630, frying pan:1311, ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>yaku (2) (have difficulty)</th>
<th>CS</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>ga</strong></td>
<td>teacher:3, government:3, person:3, ...</td>
</tr>
<tr>
<td></td>
<td><strong>wo</strong></td>
<td>hand:2950</td>
</tr>
<tr>
<td></td>
<td><strong>ni</strong></td>
<td>attack:18, action:15, son:15, ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>yaku (3) (copy; burn CDR)</th>
<th>CS</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>ga</strong></td>
<td>maker:1, distributor:1, ...</td>
</tr>
<tr>
<td></td>
<td><strong>wo</strong></td>
<td>data:178, file:107, copy:9, ...</td>
</tr>
<tr>
<td></td>
<td><strong>ni</strong></td>
<td>R:1583, CD:664, CDR:3, ...</td>
</tr>
</tbody>
</table>

...
Toyota launched the hybrid car Prius in 1997. $\Phi_1$ started selling $\Phi_2$ overseas in 2000.
Effect of Corpus Size

[ Sasano et al., NAACL-HLT09 ]

Coverage of case frames

Accuracy of case structure analysis

Accuracy of syntactic analysis

F-measure of zero anaphora resolution

Corpus size (# of sentences)
Indexing

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

Modifier → Modifiee

情報 (information) 技術の (technology) 発達は (growth) 目覚ましいものが (striking) あります (is)

進歩 (progress)

The growth of information technology is striking.

Extracted terms

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

IT 進歩 (progress)
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
  - 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%

Index

- Word sense disambiguation
  - ねぎは（syn.）
  - 風邪に効く（syn.）
  - 病気（hyp.）
  - 春（hyp.）
  - 食べ物（hyp.）

Web pages
Wikipedia

Deep NLP
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 = 0.88340

2. Facebookにおけるフィッシング対策強化を強い...  
   score = 0.87872

3. Facebookにおけるフィッシング対策強化を強い...  
   score = 0.87852

4. Facebookにおけるフィッシング対策強化を強い...  
   score = 0.851610

Googleの検索結果からも明らかだが、Facebookが直面する問題のなかで最も深刻なのは、フィッシング攻撃の問題でしょう。Facebookのアカウントを乗っ取られると、個人情報が漏洩することになるため、ユーザーの信頼を失うことになる。
**Chinese cabbage**: One of the brightly colored vegetables that plenty contains vitamin A... as an **effective vegetable** that contribute to **prevent the cold**.

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

The recovery **effect** can be obtained when eating **pumpkins** that include vitamin A. Vitamin A works to **prevent skin roughness and the cold**.
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/Contradictory Statements
Major Keywords

- "good for the environment"
- "not good for the environment"
- "CO2", "fuel consumption", "exhaust gas"

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

We can find experts on the topic!

"Japan Automobile Manufacturers Association"
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 and Hyponyms: 500K pairs

- **Dependency/Case/Ellipsis analysis**: 90% 79% F 0.40
  - 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%

- **Index**
  - word sense disambiguation
  - nezumi (hyp.)
  - hiru (syn.)
  - ooki (syn.)
  - hayashi (hyp.)
  - shoku (hyp.)

- **Web pages & Wikipedia**

- **Deep NLP**
Flow of TSUBAKI

Execute deep NLP on Azure

1. Morphological analyzer: JUMAN (C)
2. Dependency/Case/Ellipsis analyzer: KNP (C)
3. Semantic analyzer: SynGraph (Perl)

Online processing

Deep NLP

Indexing

Text data (Web pages)

Analyzed data

Index

Looking up

Search

query

Analyze search query

Doc. retrieval and ranking

Search result
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
Issues and Solutions

Our Side

- Porting Linux software to Windows
- The use of 10,000 CPU cores is challenging!
- The network between Azure (Asia) and SINET is narrow? (20 hours/3.5GB)
- Large databases of automatically acquired knowledge
- C: Cygwin (32bit)

Create a virtual hard drive (VHD) on Azure storage (Blob)

Azure Side

- Try 1x350, 2x350, 8x350
- Divide to 29x350
- Highly concurrent access to VHD on Blob
- Execute a dummy run before the real run

Manage 29 services by hand

- Strawberry Perl
- Create a manager with Azure APIs
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)
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 ...
Conclusion

Web Services

Search Engine Infra

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
  ➢ 〜10,000〜 CPUs
Future Work: Execute the Whole Process on Azure!

Offline processing

Text data (Web pages) → Deep NLP → Analyzed data → Indexing

Online processing

Search query → Search module

- Analyze search query
- Doc. retrieval and ranking

Looking up

Search result
Many thanks to

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