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

Daisuke Kawahara     Sadao Kurohashi
Kyoto University
Overview

Web Services

Search Engine

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
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, Intrasent. 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%
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. Lenovo purchased IBM in 2005 and acquired the ThinkPad brand in 2005.

ThinkPads are popular with large businesses, 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.

Contents

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
Mary ate the salad with a fork.
Mary ate the salad with mushrooms.

Knowledge Acquisition and Deep NLP

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

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
Web

15G sentences (3G pages)

Predicate-argument structures

Clustering

Case frames for 40K predicates

Parsing and Filtering

89.0% for all
98.3% for 20.7% PAs

89.0% → 89.7%

2 days

2 weeks
# Case frame examples

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

| **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, … |

...
Toyota launched the hybrid car Prius in 1997. $\Phi_1$ started selling $\Phi_2$ overseas in 2000.
Learning Curve

Accuracy of Syntactic Analysis

Accuracy of Case Structure Analysis

Coverage of Case Frames

Accuracy of Zero Anaphora Resolution

Accuracy of Synonym Recognition

Accuracy, Coverage

Corpus Size (# of sentences)

1.5M  6M  25M  100M  400M  1.6G  6.4G  15G
Knowledge acquisition and deep NLP

Ambiguous words
3000 words, 6700 word senses

Hyponyms
500K pairs

Dependency/Case/Ellipsis analysis
90% 79% F 0.40
Agent: 0.55, Intrasent. 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%

Word sense disambiguation 82.4%

Deep NLP

Index

Web pages
Wikipedia

ねぎは

風邪に

効果がある

捲宜

病気 (hyp.)

葱 (syn.)

野菜 (hyp.)

食べ物 (hyp.)

Word sense disambiguation
Architecture of Search Engine

Online process

Query \rightarrow User Interface \rightarrow Ranked Docs \rightarrow Searcher \rightarrow Query

Result \leftarrow Ranker \leftarrow Docs \leftarrow Searcher \leftarrow Terms & Condition

Offline process

Data \rightarrow Web \rightarrow Crawler \rightarrow Pages \rightarrow Indexer \rightarrow Inverted Index

Terms & Docs
Deep NLP-based Indexing

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

-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)
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)
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...
   詐欺のメールやウェブサイトに誘導されることが多く、ユーザーや管理者の注意が必要です。

2. Facebook「ユーザーデータはユーザーのもの」とっぱらり...
   Googleは長年にわたりデータプライバシーの重要な問題を提起し続けてきました。Facebookもデータプライバシーに関しては、より透明な方針を取ることが求められています。

3. Facebookに「ガラスの靴」みたいな弱点あり...
   流行のモバイルアプリの価値を高めるために、Facebookは多くの開発者にデータアクセスを許可しました。しかし、これによりセキュリティの脆弱性が生じた可能性があるため、この問題は今後も観察されるべきです。

4. Facebook - Toshiro Shimura on ...
   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いう慣習は根強く、それは多くのメンバーよりも共通されています。
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...
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.
## Comparison of Search Engines

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

<table>
<thead>
<tr>
<th>Search engine</th>
<th>MAP</th>
<th>P@3</th>
<th>P@10</th>
<th>nDCG @10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache Solr 3.6.1</td>
<td>0.124</td>
<td>0.297</td>
<td>0.276</td>
<td>0.166</td>
</tr>
<tr>
<td>TSUBAKI</td>
<td><strong>0.173</strong></td>
<td><strong>0.442</strong></td>
<td><strong>0.379</strong></td>
<td><strong>0.237</strong></td>
</tr>
</tbody>
</table>

- **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"

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

Distribution of Senders

Distribution of Opinions

Major Senders/Opinions

"Japan Automobile Manufacturers Association"

We can find experts on the topic!
Thank you!