Unsupervised and Semi-supervised learning of Nonparametric Bayesian word segmentation

Daichi Mochihashi
NTT Communication Science Laboratories, Japan
daichi@cslab.kecl.ntt.co.jp

NOE Statistical Machine Learning Seminar
2011-1-19 (Wed)
ISM
Word segmentations in NLP

- Segment string into “words”

```bash
% echo “やあこんにちは，統数研はどうですか。“
| mecab -O wakati
やあこんにちは，統数研はどうですか。
```

- Very important problem for unsegmented languages like Japanese, Chinese, Thai, Hebrew, ...
清华大学

学校概述 | 院系设置 | 管理机构 | 科学研究 | 教师队伍 | 人才培养 | 招生就业 | 虚拟校园

喜迎百年华诞 再铸新的辉煌
——2011年新年献辞

校长 王希勤

一世纪沧桑砥砺，一百年春华秋实。此时此刻，2010年的余晖散去，2011年的曙光在前！
此时此刻，清华正在送走她的第一个百年，迈进新百年的征程！在辞旧迎新、辞旧迎新之际，清华把我代表学校向全体同学、教职员工、离退休人员和海内外广大校友，向长期关心支持清华发展的各界人士，致以诚挚的敬意和新年的祝福！

一百年来，清华大学的发展始终与国家民族的命运休戚与共，形成了优良的精神传统和鲜明的办学特色、一代代清华人“自强不息、厚德载物”，涌现出众多学术大师、兴国英才和治国栋梁，为中国社会进步和世界文明发展作出了重要贡献。特别是近年来，在国家的大力支持下，学校致力于世界一流大学建设，积极探索中国特色的“大学之道”，各项事业不断取得新的进展，正在跻身世界一流大学的行列。

大学之道，育人为本。一年来，以“清华大学百年华诞的使命与挑战”为主题的第23次教育工作讨论会顺利举行。全校师生就推动高等教育优质化、培养顶尖创新人才等取得共识。

清华大学人才培养计划、以及多项教育教学改革措施相继实施。招生工作大力推动多元评价，瞄准顶尖与公平，生源质量进一步提高。就业创业中心、超前80%。到达国家重要行业和领域建功立业。

外国留学生规模不断扩大，结构进一步优化。外国研究生中非校友超过全国高校首位。

大学之道，学术为魂。一年来，我校继续面向国际学术研究前沿和国家重大战略需求开展高
Word segmentations in NLP

- Crucial first step for unsegmented languages like Japanese, Chinese, Thai, Hebrew, ...
  - Very important and fundamental problem
  - Especially: Chinese (1,200,000,000 speakers)
  - SIGHAN word segmentation Bakeoff: 2003-2010
  - Many intensive research!
Word segmentation (2)

- Learning methods so far: *Supervised*

  | S-ID:950117245-006 KNP:99/12/27 |
  | 0 5D |
  | 一方 いっぽう * 接続詞 *** |
  | 、 、 * 特殊 読点 ** |
  | 1 5D |
  | 震度 しんど * 名詞 普通名詞 ** |
  | は は * 助詞 副助詞 ** |

- “Correct” segmentations with huge human effort
  - CRF, SVM, Maxent, .. classifiers to learn

- “Correct” segmentations for speech? Twitter? Unknown language..?
  - | 女御|更衣|あ|また|さ|ぶら|ひ|た|ま|ひける|中|に|、|... |

MeCab analysis of “The Tale of Genji”, AD1000

Standard “Kyoto corpus”: 38400 hand-segmented newspaper sentences
Overview

- Unsupervised learning of word segmentation
  - HPYLM$^2$ prior + Blocked MCMC
  - “Words” for even unknown language!

- Semi-supervised learning of word segmentation
  - Systematic integration with CRF
  - Markov<->semi-Markov conversion (general)
  - MCMC-EM like algorithm for inference
Unsupervised Word Segmentation

- **Basic idea:** Maximize the probability of the segmentation $w$ of a string $s$:

$$\hat{w} = \arg\max_w p(w|s).$$

- Ex. $p(he\;sings\;a\;song) > p(hes\;ingsao\;ng)$
- *No dictionary, no “correct” supervised data*
- Find the “most natural segmentation” of a string

- **Note:** Exponential number of candidates
  - A sentence of 50 characters:
    $$2^{50} = 1,125,899,906,842,624$$ different segmentations
Probability of a sentence: \( n \)-grams

\[
p(\text{she likes music})
\]
\[
= p(\text{she}) p(\text{likes}|\text{she}) p(\text{music}|\text{likes}) p(\text{$|music})
\]

- Markov model on words: very common in NLP
  - First introduced by Shannon (1948)
- PCFG is ok too.. but simple model will suffice
  - PCFG on Twitter yellings??
- Probability tables are mostly 0
  - Hierarchical smoothing is necessary
  - Every substring could be a “word”

→ Bayesian treatment: HPYLM (yw’s talk)
Pitman-Yor n-gram model

- The Pitman-Yor (=Poisson-Dirichlet) process:
  - Draw distribution from distribution
  - Extension of Dirichlet process (w/ frequency discount)

\[ G \sim \text{PY}(G_0, \theta, d) \]

Called “base measure”

Hyperparameters (can be learned)
Hierarchical Pitman-Yor n-gram

- Kneser-Ney smoothing is an approximation of hierarchical Pitman-Yor process (Teh, ACL 2006)
  - HPYLM = “Bayesian Kneser-Ney n-gram”
Problem: Word spelling

- Possible word spelling is not uniform
  - Likely: “will”, “language”, “hierarchically”, ...
  - Unlikely: “illbe”, “nguag”, “ierarchi”, ...

- Replace the base measure using character information
  → Character HPYLM!
NPYLM: Nested Pitman-Yor LM

- Character n-gram embedded in the base measure of word n-gram
  - i.e. hierarchical Markov model
  - Poisson word length correction (see the paper)
Inference and Learning

- Simply maximize the probability of strings
  - i.e. minimize the perplexity per character of LM

- $X$: Set of strings $s_1, s_2, \cdots, s_N$
- $Z$: Set of hidden word segmentation indicators $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N$

\begin{align*}
p(X) &= \prod_{n} p(s_n) \\
p(s_n) &= \sum_{\mathbf{z}_n} p(s_n, \mathbf{z}_n) \quad \text{Hidden word segmentation of string } s_n
\end{align*}

- Notice: \textit{Exponential possibilities} of segmentations!
 Blocked Gibbs Sampling

- Sample word segmentation block-wise for each sentence (string)
  - High correlations within a sentence
Blocked Gibbs Sampling (2)

* Iteratively improve word segmentations: $\text{words}(s)$ of $s$

0. For $s = s_1 \cdots s_N$ do

   \begin{itemize}
   \item parse\_trivial($s$, $\Theta$).
   \end{itemize}

1. For $j = 1..M$ do

   \begin{itemize}
   \item For $s = \text{randperm}(s_1 \cdots s_N)$ do
   \item Remove $\text{words}(s)$ from NPYLM $\Theta$
   \item Sample $\text{words}(s) \sim p(w|s, \Theta)$
   \item Add $\text{words}(s)$ to NPYLM $\Theta$
   \item done
   \end{itemize}

   Sample all hyperparameters of $\Theta$

   done
Gibbs Sampling and word segmentation

1. 神戸では異人館街の二十棟が破損した。
2. 神戸では異人館街の二十棟が破損した。
10. 神戸では異人館街の二十棟が破損した。
50. 神戸では異人館街の二十棟が破損した。
100. 神戸では異人館街の二十棟が破損した。
200. 神戸では異人館街の二十棟が破損した。

- Iteratively resample word segmentations and update language models accordingly.
Sampling through Dynamic Programming

- Forward filtering, Backward sampling (Scott 2002)
- $\alpha[t][k]$: inside probability of substring $c_1c_2\cdots c_t$ with the last $k$ characters constituting a word
  - Recursively marginalize segments before the last $k$
Sampling through Dynamic Programming (2)

$1 \rightarrow N$  

EOS

$p(\text{EOS}|c_{N-k} \cdots c_N) \cdot \alpha[N][k]$  

- $\alpha[N][k] = \text{probability of the entire string } c_1 \cdots c_N \text{ with the last } k \text{ characters constituting a word}$
  - Sample $k$ with probability to end with EOS
- Now the final word is $c_{N-k} \cdots c_N$: use $\alpha[N-k-1][k']$ to determine the previous word, and repeat
The Case of Trigrams

In case of trigrams: use $\alpha[t][k][j]$ as an inside probability

- $\alpha[t][k][j] = \text{probability of substring with the final } k \text{ chars and the further } j \text{ chars before it being words}$
- Recurse using $\alpha[t-k-1][j][i]$ ($i = 0 \cdots L$)

>Trigrams? Practically not so necessary, but use Particle MCMC (Doucet+ 2009 to appear) if you wish
NPYLM as a Semi-Markov model

BOS ご の 東 京 都 の EOS

\[ p(\text{京都の}|\text{の東}) \]

\[ \alpha[t][k] = \sum_j p(X|Y_j)\alpha[t-k][j] \]

- Unsupervised learning of Semi-Markov HMM (Ostendorf 96, Murphy 02)
- State transition = word transition with an intensive smoothing w/ NPYLM + MCMC
English Phonetic Transcripts

- Comparison with HDP bigram (w/o character model) in Goldwater+ (ACL 2006)
- CHILDES English phonetic transcripts
  - Recover “WAtsDIs” → ”WAts DIs” (What’s this)
  - Johnson+(2009), Liang(2009) use the same data

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPY(3)</td>
<td>74.8</td>
<td>75.2</td>
<td>75.0</td>
<td>47.8</td>
<td>59.7</td>
<td>53.1</td>
</tr>
<tr>
<td>NPY(2)</td>
<td>74.8</td>
<td>76.7</td>
<td>75.7</td>
<td>57.3</td>
<td>56.6</td>
<td>57.0</td>
</tr>
<tr>
<td>HDP(2)</td>
<td>75.2</td>
<td>69.6</td>
<td>72.3</td>
<td>63.5</td>
<td>55.2</td>
<td>59.1</td>
</tr>
</tbody>
</table>

- Very small data: 9,790 sentences, 9.8 chars/sentence
Convergence & Computational time

- NPYLM is very efficient & accurate! (600x faster here)

Annealing is indispensable
## Chinese and Japanese

<table>
<thead>
<tr>
<th>Model</th>
<th>MSR</th>
<th>CITYU</th>
<th>Kyoto</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPY(2)</td>
<td>0.802 (51.9)</td>
<td>0.824 (126.5)</td>
<td>0.621 (23.1)</td>
</tr>
<tr>
<td>NPY(3)</td>
<td>0.807 (48.8)</td>
<td>0.817 (128.3)</td>
<td>0.666 (20.6)</td>
</tr>
<tr>
<td>NPY(+)</td>
<td>0.804 (38.8)</td>
<td>0.823 (126.0)</td>
<td>0.682 (19.1)</td>
</tr>
<tr>
<td>ZK08</td>
<td>0.667 (—)</td>
<td>0.692 (—)</td>
<td>—</td>
</tr>
</tbody>
</table>

- MSR&CITYU: SIGHAN Bakeoff 2005, Chinese
- Kyoto: Kyoto Corpus, Japanese
- ZK08: Best result in Zhao&Kit (IJCNLP 2008)

Note: Japanese subjective quality is much higher (proper nouns combined, suffixes segmented, etc..)
Arabic

- Arabic Gigawords 40,000 sentences (AFP news)

The Palestinians support the event because of the Islamic Resistance Movement, Hamas.

Google translate:
“Filistinebsbptazahrplansarhrkpmquompalaslamiphamas.”

Google translate:
“Palestinian supporters of the event because of the Islamic Resistance Movement, Hamas.”
first, she dreamed of little Alice herself, and once again the tiny hands were clasped upon her knee, and the bright eager eyes were looking up into hers she could hear the very tones of her voice, and see that queer little toss of her head to keep back the wandering hair that would always get into her eyes -- and still as she listened, or seemed to listen, the whole place around her became alive the strange creatures of her little sister's dream. The long grass rustled at her feet as the white rabbit hurried by -- the frightened mouse splashed his way through the neighbouring pool -- she could hear the rattle of the tea cups as the March hare and his friends shared their never-ending meal, and the shrill voice of the queen…
Conclusion of the first part

- Completely unsupervised word segmentation of arbitrary language strings
  - Combining word and character information via hierarchical Bayes
  - Very efficient using forward-backward+MCMC
- Directly optimizes Kneser-Ney language model
  - N-gram construction without any “word” information
  - Sentence probability calculation with all possible word segmentations marginalized out
    - Easily obtained from dynamic programming
Problems of Unsupervised segmentation?

- Optimize $n$-gram language models
  - Must be optimized for different tasks
    - For machine translation: Nguyen+ (COLING 2010)
    - For speech recognition: Neubig+ (Interspeech 2010)

- Not always fit for human standards
  - Ex. Inflection, proper nouns, human preference
    - “咲か”-“咲き”-“咲く”, “盧前大統領”, “そ の”
  - Remedy:
    1. Make the generative model more complex
    2. *Semi-supervised learning*
      - human standards are usually closed and general
Semi-supervised learning: JESS-CM

- JESS-CM (Suzuki+ ACL-HLT 2008): “joint probability model embedding style semi-supervised conditional model”
  - highest performance semi-supervised learning on POS tagging, NE chunking, dependency parsing

- Model:
  \[ p(y|x; \Lambda, \Theta) \propto p_{DISC}(y|x; \Lambda)p_{GEN}(y, x; \Theta)^\lambda \]
  - Product model: discriminative and generative
  - However no naïve product
    - “model weight” \(\lambda\) is included in \(\Lambda\)
    - \(\Theta\) is influenced by \(\Lambda\) through learning (i.e. recursive)
JESS-CM : inference

\[ p(y|x; \Lambda, \Theta) \propto p_{\text{DISC}}(y|x; \Lambda)p_{\text{GEN}}(y, x; \Theta)^\lambda \]

- If the discriminative model is log-linear (like CRF):
  \[ p_{\text{DISC}}(y|x) \propto \exp \left( \sum_{k=1}^{K} \lambda_k f_k(y, x) \right) \]

- Then the model is again loglinear:
  \[ p(y|x) \propto \exp \left( \lambda \log p_{\text{GEN}}(y, x) + \sum_{k=1}^{K} \lambda_k f_k(y, x) \right) \]

- Inference = maximize the objective
  \[ \log p(D|\Lambda, \Theta) = \log p(Y_s|X_s; \Lambda, \Theta) + \log p(X_u; \Lambda, \Theta) \]
  - Fix \( \Theta \) and optimize \( \Lambda \) on \( Y_s, X_s \)
  - Fix \( \Lambda \) and optimize \( \Theta \) on \( X_u \)
JESS-CM on CRF-HMM (Suzuki+ ACL08)

\[ \wedge \text{ she has a cat } \$

\text{CRF} \quad \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x) + \lambda \log p(y_t | y_{t-1}, x)

\text{HMM}

- Sum the corresponding weights of the path on the same graphical model

Interleave CRF and HMM optimization
NPYLM as a Semi-Markov model

BOS  この  東京都  の  EOS

\[
p(京都の|の東) \]

\[
\alpha[t][k] = \sum_j p(X|Y_j)\alpha[t-k][j]
\]

- NPYLM is not Markov, but semi-Markov
- State transition = word transition with an intensive smoothing w/ NPYLM + MCMC
  - CRF combination?
Semi-Markov CRF (NIPS 2004)?

- Enormous memory (1GB → 20GB)
- (Supervised) precision: at most 95%
  - Only words, no character-level information

Semi-Markov CRF is designed for only short chunks (like NE)
Learning Markov CRF ↔ Semi-Markov LM

^  この  東京  都  の  $  1

0

How to combine two different models?
- CRF → NPYLM, NPYLM → CRF
CRF $\rightarrow$ NPYLM

- Easy, proposed in (Andrew+ EMNLP 2006)
  - CRF $\rightarrow$ semi-Markov CRF
  - $p(\text{この} \rightarrow \text{東京都})$
  - Summing up the weight of features along the path
  - $\gamma(\text{start, mid, end})$
    := $\gamma(\text{start, mid}) + \gamma(\text{mid, end})$

1 = beginning of word, 0 = continuation of word
Nontrivial!

- Four patterns: $0 \rightarrow 0$, $0 \rightarrow 1$, $1 \rightarrow 0$, $1 \rightarrow 1$
- Pretend if the model is Markov HMM (not semi-Markov)
- When sentence $x$ is given, we can compute the corresponding potentials by intricately summing up NPYLM probabilities!
NPYLM→CRF (2)

Case 1→1:
1→1 = “京→都”, “京→都の”, “京→都の話”, ...

この 東 京 都 の 話

“都”

“都の”

“都の話”
NPYLM → CRF (3)

- Case 1→0:

1→0 = "東→京都", "の東→京都", "この東→京都", "東→京都の", "の東→京都の", "この東→京都の", "東→京都の話", "の東→京都の話", ⋯

<table>
<thead>
<tr>
<th></th>
<th>京都</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

This diagram illustrates the transitions and labels, with specific nodes indicating "話", "の話", and "都の話".
NPY→CRF: Code example

● C++ code for summing probabilities to compute “0→0” potential:

```c++
double sentence::ccz (int t, HPYLM *lm) {
    wstring w, h;
    int i, j, k, L = src.size();
    double z = 0;

    for (k = 0; k < MAX_LENGTH - 2; k++) {
        if (!(t + 1 + k < L)) break;
        for (j = 2 + k; j < index[t + 1 + k]; j++) {
            w = src.substr(t + 1 + k - j, j + 1);
            if (t + k - j < 0) { /* (t + 1 + k - j) - 1 */
                h = EOS;
                z += lm->ngram_probability (w, h);
            } else {
                for (i = 0; i < index[t + k - j]; i++) {
                    h = src.substr(t + k - j - i, i + 1);
                    z += lm->ngram_probability (w, h);
                }
            }
        }
    }
    return z;
}
```
What are we doing? (1)

- Mathematically, this computation is a marginalization
  - By definition,
    \[
    p(c_t^{u-1}|c_s^{t-1}) = \gamma(s, t, u) = p(z_s = 1, z_{s+1} = 0, \ldots, z_t = 1, z_{t+1} = 0, \ldots, z_u = 1)
    \]
  - Then we can marginalize:
    \[
    p(z_t = 1, z_{t+1} = 1) = \sum_k p(z_t = 1, z_{t+1} = 1, \ldots, z_k = 1)
    \]
    \[
    p(z_t = 1, z_{t+1} = 0) = \sum_l \sum_k p(z_t = 1, z_{t+1} = 0, \ldots, z_k = 1, \ldots, z_l = 1)
    \]
    \[
    p(z_t = 0, z_{t+1} = 0) = \sum_j \sum_l \sum_k p(z_{t-1} = 1, z_t = 0, z_{t+1} = 0, \ldots, z_l = 1, \ldots, z_j = 1)
    \]
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of $1 \rightarrow 1$:
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of $0 \rightarrow 1$:
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of $1 \rightarrow 0$:
What are we doing? (2)

*Graphically*, summing the collection of paths for the case of $0 \rightarrow 0$:
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of $0\to0$:
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of $0 \rightarrow 0$:
What are we doing? (2)

- Graphically, divide the paths at the section into four bins:

  Relevant area, crossing the section
Experiments (still ongoing)

- Sina Microblog (新浪微博)
  - Chinese equivalent of Twitter, 94,800,000 users
  - Full of non-standard Chinese sentences

- Japanese blog (Shokotan blog)
  - Famous for being full of jargons, new proper nouns, ..

- CSJ (Corpus of Spontaneous Japanese)
  - by National Institute of Japanese language and linguistics, next to ISM

- SIGHAN word segmentation Bakeoff 2005
  - Public dataset, intensively studied, newspaper

Tremendous!
中 三 の とき 後楽園 遊園地 に タイムレンジャーショー を 見 に 行き
まくってた ころ の こと そう だ お ね、 セル画 は 修正 に 全て 塗り 直し
とかある だろう けど デジタル なら 一 発 で レイヤー 直せる から ... 戸田 さんの 生 歌声 が 最高に 美しくて チャーム 狀態 にな り ました。 そして ザ・ピーナッツ 役の 堀内 敬子 さん 瀬戸 カトリーヌ さんが 息
ピッタリ に 渡辺 プロ の そうそうたる 名曲 を 歌いあげ、 最高 の ハーモ
ニー で とにかく すばらしかった です。 生歌 で あの 美しさ ... 。 四 つ と も 全部 この 川柳 wwwwww お茶 wwwww イト カワユス
wwwww イト カワユス wwwww (^ω^)(^ω^)(^ω^)
深夜 まで お疲れ さ マミタス(°’ω’°) ギャル 曽根 たん ！ 最近 よく
一緒 になる の と 楽屋 に 遊び に きて くれる の で いろいろ おしゃべり
して タノジス ！ (^ω^) 今 日 も いろいろ 話し た お ね
イプサ の 化粧 水 が ケア 楽 チン だ し 肌 に ギザ あう ！ これ ギガント
肌調子 よく なり まんた (^ω^)

● Supervised : Kyoto corpus 37,400 setences
Unsupervised: Shokotan blog 40,000 sentences
# Shokotan blog: words

- **Excerpt of random words with frequency >= 2**

<table>
<thead>
<tr>
<th>word</th>
<th>frequency</th>
<th>word</th>
<th>frequency</th>
<th>word</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>あるるるる</td>
<td>2</td>
<td>スワロフスキー</td>
<td>3</td>
<td>早いし</td>
<td>2</td>
</tr>
<tr>
<td>ますえ</td>
<td>2</td>
<td>わたる</td>
<td>11</td>
<td>信じろ</td>
<td>6</td>
</tr>
<tr>
<td>そびれちゃった</td>
<td>2</td>
<td>コマ送り</td>
<td>3</td>
<td>似てる</td>
<td>26</td>
</tr>
<tr>
<td>メリクリスマース</td>
<td>3</td>
<td>おおっお</td>
<td>7</td>
<td>居る</td>
<td>10</td>
</tr>
<tr>
<td>シクシク</td>
<td>3</td>
<td>にじむ</td>
<td>4</td>
<td>よる</td>
<td>85</td>
</tr>
<tr>
<td>チーム</td>
<td>45</td>
<td>箔</td>
<td>12</td>
<td>LaQua</td>
<td>7</td>
</tr>
<tr>
<td>ロック</td>
<td>11</td>
<td>ギギ</td>
<td>2</td>
<td>ただただ</td>
<td>7</td>
</tr>
<tr>
<td>キムタク</td>
<td>12</td>
<td>呼んで</td>
<td>29</td>
<td>ストロベリメロディ</td>
<td>21</td>
</tr>
<tr>
<td>うなぁ</td>
<td>2</td>
<td>席</td>
<td>31</td>
<td>スターサーサートゥハッッッ</td>
<td>2</td>
</tr>
<tr>
<td>したろう</td>
<td>3</td>
<td>100</td>
<td>55</td>
<td>ひろがって</td>
<td>3</td>
</tr>
<tr>
<td>去った</td>
<td>4</td>
<td>グラビア</td>
<td>85</td>
<td>しろま</td>
<td>3</td>
</tr>
<tr>
<td>死兆星</td>
<td>4</td>
<td>田尻</td>
<td>3</td>
<td>カワユスピンク</td>
<td>2</td>
</tr>
<tr>
<td>スッキリ</td>
<td>6</td>
<td>より焼き</td>
<td>2</td>
<td>海馬</td>
<td>3</td>
</tr>
<tr>
<td>ドバアア</td>
<td>2</td>
<td>ヒヤダルコ</td>
<td>3</td>
<td>除外</td>
<td>3</td>
</tr>
<tr>
<td>開催</td>
<td>47</td>
<td>永久</td>
<td>34</td>
<td>けえ</td>
<td>6</td>
</tr>
<tr>
<td>おく</td>
<td>17</td>
<td>ヤマト</td>
<td>2</td>
<td>なんとゆう</td>
<td>2</td>
</tr>
</tbody>
</table>
Sina microblog (新浪微博)

今天一大早就被电话吵醒了，折磨死我了，昨天太晚睡了，早上被这电话搞得晕忽忽！
头疼，发热。。。貌似感冒了，晚上睡觉不能裸睡了。要穿睡衣了。

咿~ ? 半个钟前发的围脖咋不见了咧~~ 只是感慨了一下今天的归途特顺嘛~~~ (╯﹏╰)b

下雨了，不知道广州那边有没有下雨，明天的同学聚会我去不了了，[伤心]大哭

学校附近一隻很可爱的狗狗，做了點特效[心][心][心]我們學校學生超愛牠的！！！！[哈哈]
明儿我要把中山陵搞定~~~~~ 玛丽隔壁的~~~(-_-)
好饿啊...走！妈妈带你出去吃饭去～～～～(((((CTORो=.^·ェ·)))

喵~ o(∩ω∩=)m

梦。。。混乱的梦。。清晰的梦。。。。。。

● Supervised ：MSR 87000 sentences (Mandarin)
  Unsupervised ：Sina API, 98700 sentences
Corpus of Spontaneous Japanese (CSJ)

- **Experimental setting**
  - Supervised: Kyoto corpus+CSJ 10% (21000 sents)
  - Unsupervised: CSJ 90% (190000 sents)
  - Training: CRF (only supervised) / NPYLM (sup+unsup)

- **Results w.r.t. CSJ hand segmentations**
  - $P/R/F=0.928/0.893/0.91\rightarrow0.931/0.899/0.915$
  - Generally better, but occasionally gets worse:
    - CRF: ええだから何て言うんでしょうか立ってるとお腹に力..
    - NPY: ええだから何て言うんでしょうか立ってるとお腹に力..
    - CRF: 神奈川寄りとかあっちの方行っちゃいますよね値段..
    - NPY: 神奈川寄りとかあっちの方行っちゃいますよね値段..
    - CRF: 力強い方向性あるタッチ
    - NPY: 力強い方向性あるタッチ
SIGHAN Bakeoff 2005

- Publicly available dataset of Chinese word segmentation, MSR portion
  - Mostly newspaper, *not so much suited for our task but standard*

- Data: MSR supervised (87000 sentences) + Chinese Gigaword (200000 sentences)

- Results:
  - Baseline: 97.4% F-measure with features of (Xu+ 2009)
  - MSR+Gigaword: 97.5%
  - MSR+Gigaword+dict: 97.5%

World best baseline! (closed)

97.3% on DPLVM

Difficult to beat.. Much more data to cover test data (5 days on Xeon W5590 to compute)
Conclusion

- Semi-supervised learning of CRF-NPYLM
  - Product models, each depends each
- Convert Semi-Markov $\leftrightarrow$ Markov models
  - General observation, not only to NPYLM
- Good for blogs and Twitters
  - Maintaining accuracy on supervised data
  - Need: huge unsupervised data to improve on standard datasets
    - Parallelization or other implementation techniques
Future work

- Predicting POS (part-of-speech) simultaneously
  - “Class-based” n-gram models + Multiclass CRF
  - Class-based language models will also improve on unsupervised data
    - Ex. Predict “ギザ カワユス” → “ギザ スゴス”
- However, in the current NLP...
  - No standard theory for building class-based n-grams. only heuristic (like Google Mozc)
  - Dealing with “hierarchy” of part-of-speech
    - Ex. Noun → Proper noun → Name of a person
    - Theory of Markov model on tree-structured classes? (imagine hLDA)