Brown et al. 1992 Clustering

Simon Šuster
University of Groningen

24. 5. 2013

Background

- Introduced by **Brown, Della Pietra, deSouza, Lai and Mercer** in 1992 [Brown et al., 1992]
- Referred to as "Brown clustering" (rarely, IBM clustering)
- Brown was fortunate . . . (cf. [Metropolis et al., 1953])

- Relatively simple algorithm
- Popular, cited in 345 papers (ACM DL)

General*

- **Idea**: partition vocabulary in the corpus to clusters

- **Input**: raw (or tokenized) text
- **Output**: clusters, hierarchical

- Clusters (ideally) include semantically similar words
- No supervision necessary

*This slide adapted from [Collins, 2011]

General procedure

1. start with vocabulary $\mathcal{V}$
2. initialize: put $\mathcal{V}$ into distinct[1] clusters $\Rightarrow$ obtain clustering $\mathcal{C}$
3. iteratively merge[2] two[3] clusters that maximize Quality($\mathcal{C}$)

[†] With some table-keeping of ΔQuality($\mathcal{C}$)

General procedure

1. start with vocabulary $\mathcal{V}$
2. initialize: put $\mathcal{V}$ into distinct[1] clusters $\Rightarrow$ obtain clustering $\mathcal{C}$
3. iteratively merge[2] two[3] clusters that maximize Quality($\mathcal{C}$)

- Note
  1. hard clustering
  2. agglomerative: tree structure
  3. binary tree

- Runs in $O(|\mathcal{V}|^3)^\dagger$

†With some table-keeping of ΔQuality($\mathcal{C}$)

Optimized variant

Idea: restrict n of clusters to k

1. initialize:
   - sort $\mathcal{V}$ by freq
   - put first k types into distinct clusters $\Rightarrow$ again, obtain $\mathcal{C}$
2. iterate:
   - put $k+1^{st}$ type to a new cluster
   - merge the pair in $k+1$ clusters that maximizes Quality($\mathcal{C}$)
3. iteratively merge the remaining k clusters (build tree as previously)

- Runs in O($k^2|\mathcal{V}|$)

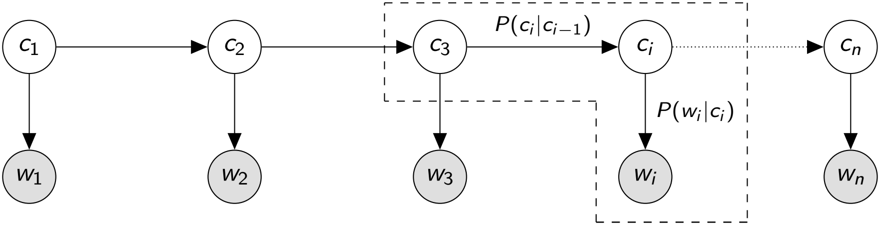
What is **Quality($\mathcal{C}$)**?

Quality($\mathcal{C}$)

- Context: class-based bigram language model

$$
\begin{aligned}
Quality(\mathcal{C}) &= \frac{1}{n} log P(w_1, \ldots, w_n) \\
&\hspace{10em} \text{(function of probability of a sequence)} \\
&= \frac{1}{n} log P(w_1, \ldots, w_n, C(w_1), \ldots, C(w_n)) \\
&\hspace{10em} \text{(expand with deterministic mapping)} \\
&= \frac{1}{n} log \prod_{i=1}^{n} \underbrace{P(C(w_i)|C(w_{i-1}))}_{\text{transition prob.}} \underbrace{P(w_i|C(w_i))}_{\text{emission prob.}} \qquad \text{(model)}
\end{aligned}
$$

Model as a Bayesian network



Quality($\mathcal{C}$) cont'd

To repeat:

$$Quality(\mathcal{C}) = \frac{1}{n} log \prod_{i=1}^{n} P(C(w_i)|C(w_{i-1}))P(w_i|C(w_i)) \qquad \text{(model)}$$

decomposes to ...

$$= \sum_{c,c'} P(c,c') log \frac{P(c,c')}{P(c)P(c')} + \sum_{w} P(w) log P(w)$$

$$= I(\mathcal{C}) - H$$

- Entropy H is constant
- Mutual information $I(\mathcal{C})$ defines Quality($\mathcal{C}$)!

- Let table L keep track of change in Quality
- Merge clusters m, n[‡] having the maximum score (least decrease) in L

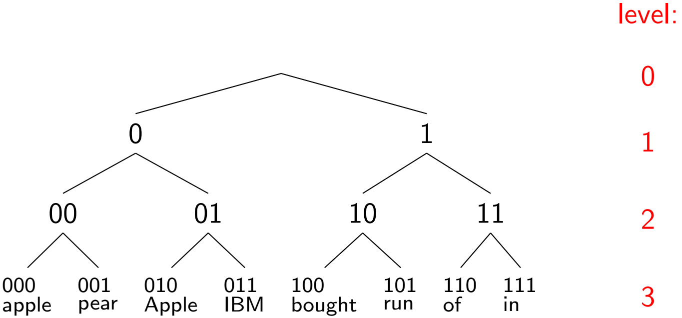
$$L(m,n) = \sum_{d \in \mathcal{C}'} I(m \cup n, d) - \sum_{d \in \mathcal{C}} (I(m,d) + I(n,d)),$$

where:
$m \cup n =$ the new cluster
$\mathcal{C} =$ the current set of clusters
$\mathcal{C}' = \mathcal{C} - \{m,n\} + \{m \cup n\}$ the set of clusters after merging m, n
$I =$ MI weight between two adjacent clusters

[‡] whichever, regardless of adjacency

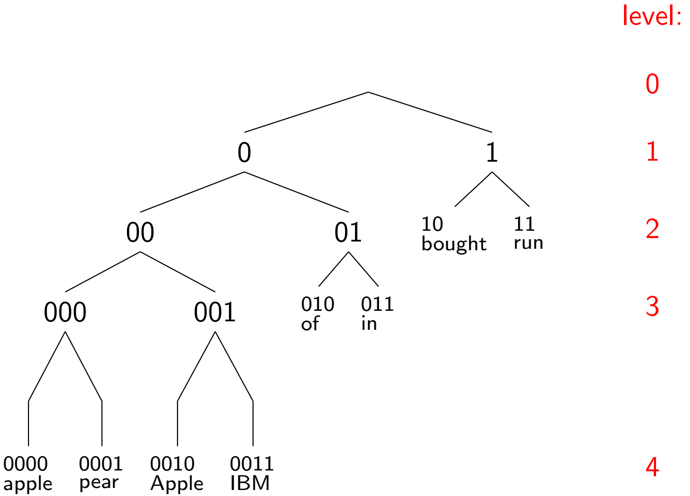
Illustration

- A perfect balanced binary tree



Illustration

- In reality:



- Not balanced: minor consequence on filtering by prefix

Example clusters from Brown et al. 1992

Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays

people guys folks fellows CEOs chaps doubters commies unfortunates blokes

down backwards ashore sideways southward northward overboard aloft downwards adrift

water gas coal liquid acid sand carbon steam shale iron

great big vast sudden mere sheer gigantic lifelong scant colossal

American Indian European Japanese German African Catholic Israeli Italian Arab

mother wife father son husband brother daughter sister boss uncle

machine device controller processor CPU printer spindle subsystem compiler plotter

John George James Bob Robert Paul William Jim David Mike

feet miles pounds degrees inches barrels tons acres meters bytes

had hadn't hath would've could've should've must've might've

that tha theat

head body hands eyes voice arm seat eye hair mouth

Clusters for Dutch

- Percy Liang's implementation in C++ [Liang, 2005]
- **SoNaR**: random sample of 4M sents, tokenized
- remove sents of length $\leq 4 \Rightarrow$ 46M tokens
- remove words with freq $< 3 \Rightarrow$ **288k** types

- **1000** clusters
- 95 hours, single core (i5 2.67GHz)

Clusters for Dutch: some statistics

Population, n of clusters = 1000

Min.	Median	Mean	Max.
2	97	288.1	16660

Example clusters from Dutch SoNaR (46M)

vrijdagavond woensdag nieuwjaarsdag woensdagvoormiddag di. Koningsdag +120 others

Tandarts Ceo Minister Coach Wereldkampioen Columnist Gastvrouw Frontman +1190

zijdelings vanbinnen rechtstaand daarbuiten achterin overdag ergens +74

vaak regelmatig zelden nimmer uitdrukkelijk sporadisch normalerwijze +18

hem 'm + 40

Clerck Clercq Vries Vos Haan Mulder Villepin + 1900

Spa Fra Ita belga EEG + 1285

prijs koers rente score balans marge + 692

conservatief mager dun klein piepklein statisch idyllisch sappig getalenteerd +585

dàt dat dát datje dan +10

behoeft wenste durfde hoefde wenst hoeft durft +16

Example raw output from SoNaR

...
10111100110 biertjes 29
10111100110 jaartjes 39
10111100110 ogenblikken 105
10111100110 zondagen 117
10111100110 druppels 146
10111100110 uurtjes 208
10111100110 werkdagen 239
10111100110 nachten 549
10111100110 eeuwen 815
10111100110 uren 2436
10111100110 dagen 14479
...
10111100111 innings 16
10111100111 kalenderjaren 17
10111100111 legislaturen 19
10111100111 setballen 20
10111100111 kwartalen 126
10111100111 decennia 907
10111100111 seizoenen 1142
10111100111 weken 11322
10111100111 maanden 14513

Some applications

- Dependency parsing [Koo et al., 2008, Haffari et al., 2011] (inter alia)
- PCFG parsing [Candito and Crabbé, 2009]
- Semantic dependency parsing [Zhao et al., 2009]
- Named-entity recognition [Turian et al., 2010, Miller et al., 2004]
- QA [Momtazi and Klakow, 2009]

Extension of the Brown algorithm (exchange algorithm)

- [Martin et al., 1998]
- [Uszkoreit and Brants, 2008]

Critical view

- Insensitive to underlying sentence structure
- Hard clustering and sense conflation
- Time complexity
- Local optima (greedy merging)

What do {kleding, afkomst, humor, infrastructuur, software, poëzie, landbouw, wijn} have in common?

Bibliography I

Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992).
Class-based n-gram models of natural language.
Computational Linguistics, 18(4):467–479.

Candito, M. and Crabbé, B. (2009).
Improving generative statistical parsing with semi-supervised word clustering.
In *Proceedings of the 11th International Conference on Parsing Technologies*, IWPT '09, pages 138–141.

Collins, M. (2011).
The Brown et al. Word Clustering Algorithm. Presentation.
`http://www.cs.columbia.edu/~cs4705/fall2011/lectures/brown.pdf`.

Haffari, G., Razavi, M., and Sarkar, A. (2011).
An Ensemble Model that Combines Syntactic and Semantic Clustering for Discriminative Dependency Parsing.
In *ACL*, pages 710–714.

Bibliography II

Koo, T., Carreras, X., and Collins, M. (2008).
Simple Semi-supervised Dependency Parsing.
In *Proceedings of ACL-08: HLT*, pages 595–603, Columbus, Ohio. ACL.

Liang, P. (2005).
Semi-supervised learning for natural language.
Master's thesis, Massachusetts Institute of Technology.

Martin, S., Liermann, J., and Ney, H. (1998).
Algorithms for bigram and trigram word clustering.
Speech Communication, 24(1):19–37.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953).
Equation of State Calculations by Fast Computing Machines.
The Journal of Chemical Physics, 21(6):1087–1092.

Miller, S., Guinness, J., and Zamanian, A. (2004).
Name tagging with word clusters and discriminative training.
In *HLT-NAACL*, pages 337–342.

Bibliography III

Momtazi, S. and Klakow, D. (2009).
A word clustering approach for language model-based sentence retrieval in question answering systems.
In *CIKM*, pages 1911–1914.

Turian, J., Ratinov, L., and Bengio, Y. (2010).
Word representations: a simple and general method for semi-supervised learning.
ACL '10, pages 384–394.

Uszkoreit, J. and Brants, T. (2008).
Distributed word clustering for large scale class-based language modeling in machine translation.
In *ACL*, pages 755–762.

Zhao, H., Chen, W., Kit, C., and Zhou, G. (2009).
Multilingual dependency learning: a huge feature engineering method to semantic dependency parsing.
CoNLL '09, pages 55–60.