Enhanced Brown clustering with dependencies

Simon Šuster and Gertjan van Noord
University of Groningen

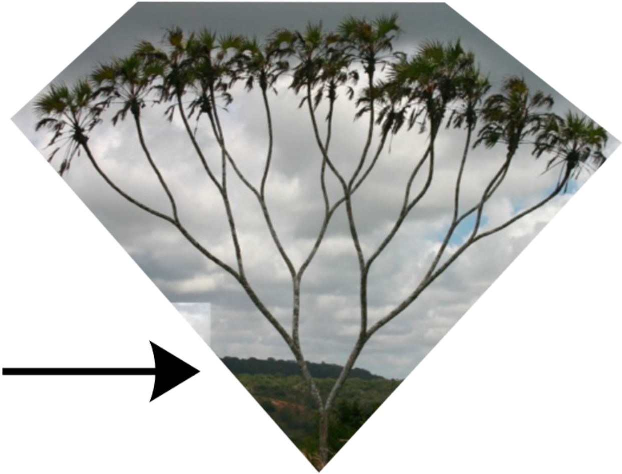
CLIN, January 10, 2014

Brown clustering

- Grouping similar words (semantic; paradigmatic & orthographic variants)
- Extensively used in NLP additional features in NER, parsing, question answering etc.
- Addresses lexical sparseness
- Robust
- No vectorization or feature design needed

Clustering intuition

Words/text



Word clusters in a binary tree

Clustering procedure

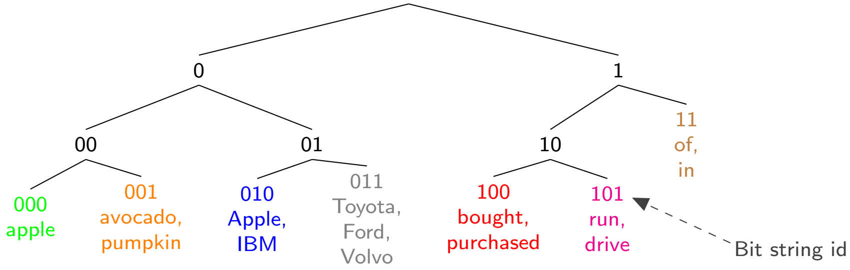
(Simplified:)
k=number of clusters

- put k most frequent words into k distinct clusters
- merge remaining words with the existing k clusters, one by one
  - (words now grouped, no hierarchy yet)
- merge clusters to build a binary tree, bottom-up

Clustering procedure

(Simplified:)
k=number of clusters

- put k most frequent words into k distinct clusters
- merge remaining words with the existing k clusters, one by one
    - (words now grouped, no hierarchy yet)
- merge clusters to build a binary tree, bottom-up

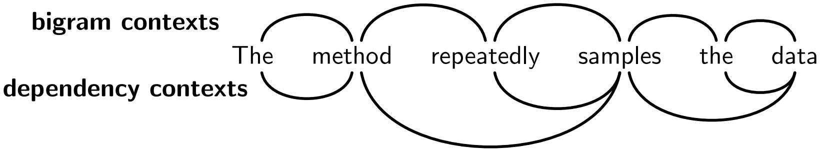


Introducing dependencies I

- "Merge": minimizing the loss in average mutual information between clusters
- MI is derived from a class-based bigram language model
  - Word class conditioning on the class of the *previous* word
- Local-only representation is a **limitation**

Idea:

- Establish context with dependencies
  (assuming we can trust the parser...)



Introducing dependencies II

- Paraphrase the model with a dependency language model
- Using a simple factorization: words conditioned on their heads

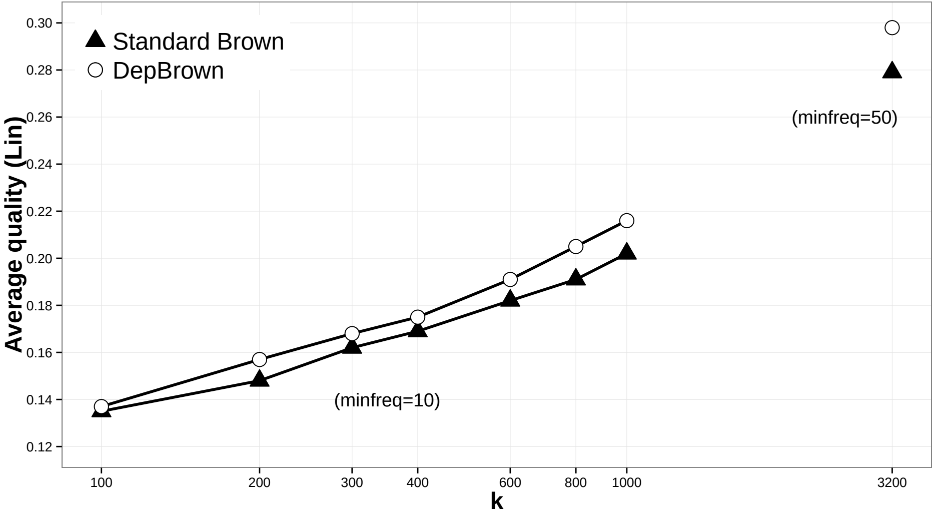
Concretely:

- Modify the code by P. Liang slightly
- Parse a 46M-word sample from SoNaR with Alpino parser
- Feed the dependency instances to the clustering software
- Evaluate: wordnet similarity task (Cornetto)
  - Average similarity over all clusters, as measured by Lin score

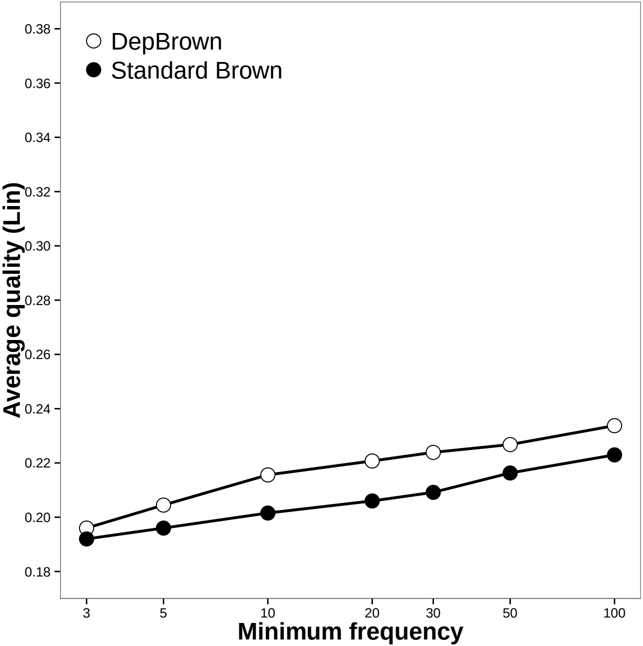
Group	Cluster id	Most frequent words
A1	001010001011100	aannemer, huis_arts, bakker, notaris, apotheker, makelaar, projectontwikkelaar, postbode,...
A2	001010001011011	analist, criticus, waarnemer, kenner, commentator, mens_recht_organisatie, insider,...
A3	0010100010111110	ondernemer, zakenman, bedrijf_leider, zelfstandige, koopman, starter, ambachtsman,...
B1	011101111011110	mij
B2	01110111101110	zichzelf, mezelf, jezelf, onszelf, mijzelf, uzelf
B3	01110111101100	hen
C1	00110010010	Bush, Obama, Clinton, Poetin, Chirac, Sarkozy,...
C2	0011000111010	Sarah, Kim, Nathalie, Justine, ...
C3	0011000111011	David, Jimmy, Benjamin, ...
D1	001011100010101	email, mail, sms, sms_DIM, e-mail, mail_DIM, ...
D2	001011100010100	telefoon, satelliet, telefonie, telefoon_lijn, Explorer, muziek_speler, iTunes,...
E	001000010110101	inkomen, energie_verbruik, minimum_loon, cholesterol, opleidingsniveau, IQ, alcohol_gehalte,...

Group	Cluster id	Most frequent words
A1	001010001011100	contractor, family doctor, baker, lawyer, pharmacist, real estate agent, property developer, postman, ...
A2	001010001011011	analyst, reviewer, observer, expert, commentator, people's rights organisation, insider, ...
A3	0010100010111110	entrepreneur, businessman, manager, self-employed, merchant, starter, craftsman, ...
B1	011101111011110	me
B2	01110111101110	him/herself, myself, yourself
B3	01110111101100	them
C1	00110010010	Bush, Obama, Clinton, Putin, ...
C2	0011000111010	Sarah, Kim, Nathalie, Justine, ...
C3	0011000111011	David, Jimmy, Benjamin, ...
D1	001011100010101	email, mail, sms, sms_DIM, e-mail, mail_DIM, ...
D2	001011100010100	telephone, satellite, telephony, telephone line, Explorer, music player, iTunes, ...
E	001000010110101	income, energy consumption, minimum wage, cholesterol, IQ, alcohol content, ...

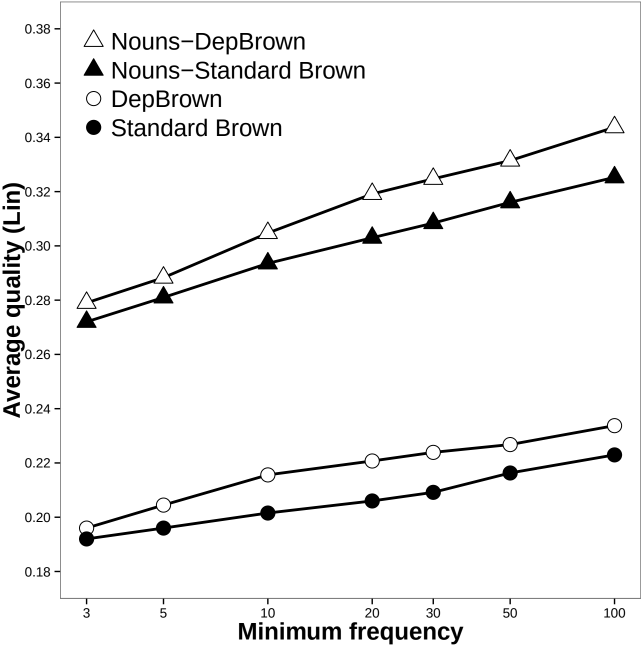
Varying k number of clusters



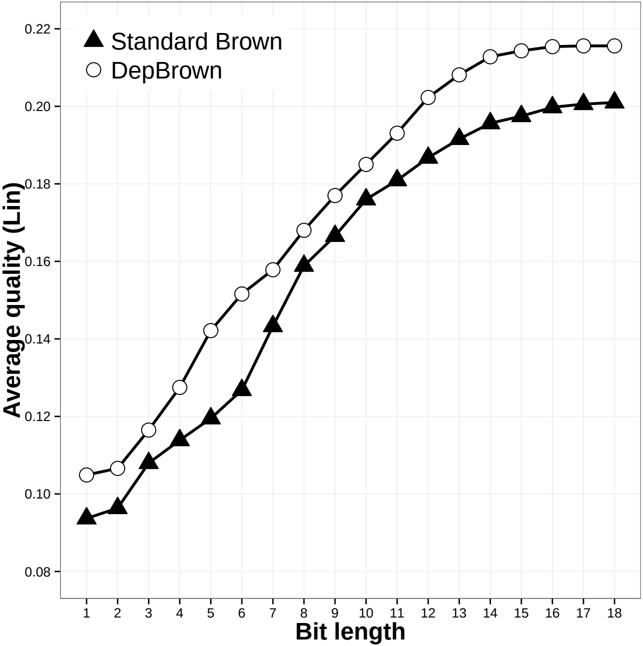
Varying *minfreq*



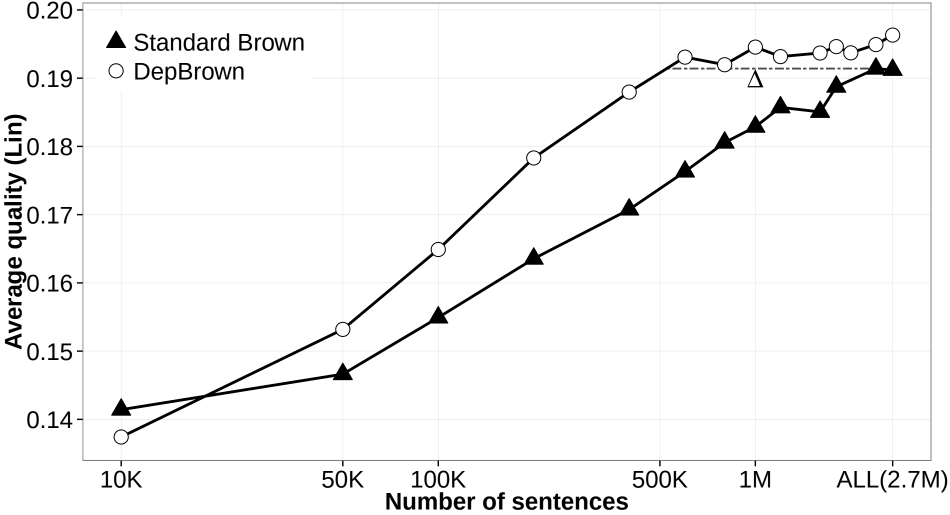
Varying *minfreq* + Nouns only



Prefix length



Amount of data



Using *labeled* dependencies

Dependency relation selection:

- clustering instances belonging to a specific relation (45 r.)
- **better** than unlabeled-dependency clustering from before:
  - subjects
  - direct objects
  - directional complements
  - 2-nd order (intervening preposition) dir. & prep. complements

Labeled dependencies: next steps

- Selection
  - Determines the input text for clustering
  - Idea: some relations yield less syn/sem coherent clusters
  - Drawback: at clustering time, no differentiation made between relations
- Separate modeling
  - Different contexts contribute differently
  - When clustering e.g. a verb, distinguish between SU and OBJ relations
  - Explicitly mark words with relations
  - Or reformulate the model

This talk

- Brown clustering intuition and procedure
- Alternative view: dependencies
- Similarity task evaluation
- Encouraging results for dependency clustering
  - k number of clusters
  - minimum frequency
  - nouns only
  - prefix length
  - learning curve
  - labeled dependencies