

Grouping Words

PA154 Language Modeling (8.1)

Pavel Rychlý

pary@fi.muni.cz

April 16, 2024

Source: Natural Language Processing (600.465)
Jason Eisner, Johns Hopkins Univ.
www.cs.jhu.edu/~jason/465/

Pavel Rychlý • Grouping Words • April 16, 2024

1 / 27

Linguistic Objects in this Course

- **Trees (with strings at the nodes)**
 - Syntax, semantics
 - Algorithms: *Generation, parsing, inside-outside, build semantics*
- **Sequences (of strings)**
 - n-grams, tag sequences
 - morpheme sequences, phoneme sequences
 - Algorithms: *Finite-state, best-paths, forward-backward*
- **"Atoms" (unanalyzed strings)**
 - Words, morphemes
 - Represent by contexts – other words they occur with
 - Algorithms: *Grouping similar words, splitting words into senses*

Pavel Rychlý • Grouping Words • April 16, 2024

2 / 27

A Concordance for "party" - from www.webcorp.org.uk

WebCorp Live
Concordance the web in real-time.

Search Wordlist Tool User Guide WebCorp LSE Publications Feedback

WebCorp Live lets you access the Web as a corpus - a large collection of texts from which examples of real language use can be extracted. [More...](#)

Search:

Case Insensitive: Span:

Search API: Language:

[Advanced Options](#)

Pavel Rychlý • Grouping Words • April 16, 2024

3 / 27

A Concordance for "party" - from www.webcorp.org.uk

- thing. She was talking at a **party** thrown at Daphne's restaurant in
- have turned it into the hot dinner-**party** topic. The comedy is the
- selection for the World Cup **party**, which will be announced on May 1
- in the 1983 general election for a **party** which, when it could not bear to
- to attack the Scottish National **Party**, who look set to seize Perth and
- that had been passed to a second **party** who made a financial decision
- the by-pass there will be a street **party**. "Then," he says, "we are going
- number-crunchers within the Labour **party**, there now seems little doubt
- political tradition and the same **party**. They are both relatively Anglophilic
- he told Tony Blair's modernised **party** they must not retreat into "warm
- "Oh no, I'm just here for the **party**," they said. "I think it's terrible
- A future obliges each **party** to the contract to fulfil it by
- be signed by or on behalf of each **party** to the contract." Mr David N

Pavel Rychlý • Grouping Words • April 16, 2024

4 / 27

What Good are Word Senses?

- thing. She was talking at a **party** thrown at Daphne's restaurant in
- have turned it into the hot dinner-**party** topic. The comedy is the
- selection for the World Cup **party**, which will be announced on May 1
- in the 1983 general election for a **party** which, when it could not bear to
- to attack the Scottish National **Party**, who look set to seize Perth and
- that had been passed to a second **party** who made a financial decision
- the by-pass there will be a street **party**. "Then," he says, "we are going
- number-crunchers within the Labour **party**, there now seems little doubt
- political tradition and the same **party**. They are both relatively Anglophilic
- he told Tony Blair's modernised **party** they must not retreat into "warm
- "Oh no, I'm just here for the **party**," they said. "I think it's terrible
- A future obliges each **party** to the contract to fulfil it by
- be signed by or on behalf of each **party** to the contract." Mr David N

Pavel Rychlý • Grouping Words • April 16, 2024

5 / 27

What Good are Word Senses?

- thing. She was talking at a **party** thrown at Daphne's restaurant in
- have turned it into the hot dinner-**party** topic. The comedy is the
- selection for the World Cup **party**, which will be announced on May 1
- the by-pass there will be a street **party**. "Then," he says, "we are going
- "Oh no, I'm just here for the **party**," they said. "I think it's terrible
- in the 1983 general election for a **party** which, when it could not bear to
- to attack the Scottish National **Party**, who look set to seize Perth and
- number-crunchers within the Labour **party**, there now seems little doubt
- political tradition and the same **party**. They are both relatively Anglophilic
- he told Tony Blair's modernised **party** they must not retreat into "warm
- that had been passed to a second **party** who made a financial decision
- A future obliges each **party** to the contract to fulfil it by
- be signed by or on behalf of each **party** to the contract." Mr David N

Pavel Rychlý • Grouping Words • April 16, 2024

6 / 27

What Good are Word Senses?

John threw a "rain forest" party last December. His living room was full of plants and his box was playing Brazilian music ...

What Good are Word Senses?

- Replace word **w** with sense **s**
 1. *Splits w* into senses: distinguishes this token of **w** from tokens with sense **t**
 2. *Groups w* with other words: groups this token of **w** with tokens of **x** that also have sense **s**

What Good are Word Senses?

- number-crunchers within the Labour **party**, there now seems little doubt
- political tradition and the same **party**. They are both relatively Anglophilic
- he told Tony Blair's modernised **party** they must not retreat into "warm
- thing. She was talking at a **party** thrown at Daphne's restaurant in
- have turned it into the hot dinner-**party** topic. The comedy is the
- selection for the World Cup **Party**, which will be announced on May 1
- the by-pass there will be a street **party**. "Then," he says, "we are going
- "Oh no, I'm just here for the **party**," they said. "I think it's terrible
- an appearance at the annual awards **bash**, but feels in no fit state to
- -known families at a fundraising **bash** on Thursday night for Learning
- Who was paying for the **bash**? The only clue was the name Asprey,
- Mail, always hosted the annual **bash** for the Scottish Labour front-
- popular. Their method is to **bash** sense into criminals with a short,
- just cut off people's heads and **bash** their brains out over the floor,

What Good are Word Senses?

- number-crunchers within the Labour **party**, there now seems little doubt
- political tradition and the same **party**. They are both relatively Anglophilic
- he told Tony Blair's modernised **party** they must not retreat into "warm
- thing. She was talking at a **party** thrown at Daphne's restaurant in
- have turned it into the hot dinner-**party** topic. The comedy is the
- selection for the World Cup **party**, which will be announced on May 1
- the by-pass there will be a street **party**. "Then," he says, "we are going
- "Oh no, I'm just here for the **party**," they said. "I think it's terrible
- an appearance at the annual awards **bash**, but feels in no fit state to
- -known families at a fundraising **bash** on Thursday night for Learning
- Who was paying for the **bash**? The only clue was the name Asprey,
- Mail, always hosted the annual **bash** for the Scottish Labour front-
- popular. Their method is to **bash** sense into criminals with a short,
- just cut off people's heads and **bash** their brains out over the floor,

What Good are Word Senses?

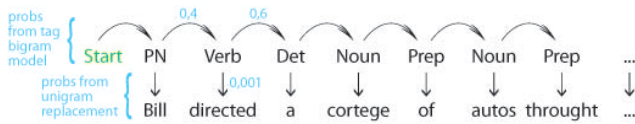
- **Semantics / Text understanding**
 - Axioms about TRANSFER apply to (some tokens of) throw
 - Axioms about BUILDING apply to (some tokens of) bank
- **Machine translation**
- **Info retrieval / Question answering / Text categ.**
 - Query or pattern might not match document exactly
- **Backoff for just about anything**
 - what word comes next? (speech recognition, language ID,...)
 - trigrams are sparse but tri-meanings might not be
 - billexical PCFGs:
 - $p(S[\text{devour}] \rightarrow NP[\text{lion}] VP[\text{devour}] | S[\text{devour}])$
 - approximate by $p(S[\text{EAT}] \rightarrow NP[\text{lion}] VP[\text{EAT}] | S[\text{EAT}])$
- **Speaker's real intention is senses; words are a noisy channel**

Cues to Word Sense

- Adjacent words (or their senses)
- Grammatically related words (subject, object, ...)
- Other nearby words
- Topic of document
- Sense of other tokens of the word in the same document

Word Classes by Tagging

- Every tag is a kind of class
- Tagger assigns a class to each word token



Word Classes by Tagging

- Every tag is a kind of class
- Tagger assigns a class to each word token
 - Simultaneously groups and splits words
 - “party” gets split into N and V senses
 - “bash” gets split into N and V senses
 - {party/N, bash/N} vs. {party/V, bash/V}
 - What good are these groupings?

Learning Word Classes

- Every tag is a kind of class
- Tagger assigns a class to each word token
 - {party/N, bash/N} vs. {party/V, bash/V}
 - What good are these groupings?
 - Good for predicting next word or its class!
- Role of forward-backward algorithm?
 - It adjusts classes etc. in order to predict sequence of words better (with lower perplexity)

Words and Vectors

- Represent each word **type** w (party) by a point in k -dimensional space
 - e.g., k is size of vocabulary
 - the 17^{th} coordinate of w represents **strength** of w 's association with vocabulary word 17

Word	Count	
aardvark	0	
abacus	0	
abandoned	3	
abbot	1	
abduct	0	
above	7	too high
...		
zygote	1	
zymurgy	0	too low

From corpus:

Jim Jeffords **abandoned** the Republican **party**.
 There were lots of **abbots** and nuns dancing at that **party**.
 The **party above** the art gallery was, **above** all, a laboratory for synthesizing **zygotes** and beer.

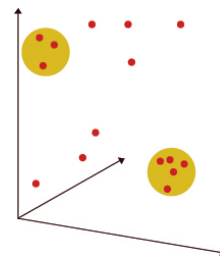
- Represent each word **type** w (party) by a point in k -dimensional space
 - e.g., k is size of vocabulary
 - the 17^{th} coordinate of w represents **strength of w 's association** with vocabulary word 17.
- *How might you measure this?*
 - how often words appear next to each other
 - how often words appear near each other
 - how often words are syntactically linked
 - should correct for commonness of word (e.g., “above”)

- Represent **each word type w (party)** by a point in k -dimensional space
 - e.g., k is size of vocabulary
 - the 17th coordinate of w represents **strength** of w 's association with vocabulary word 17.
- Plot all word types in k -dimensional space
- Look for **clusters** of close-together types

Learning Classes by Clustering

- Plot all word types in k -dimensional space
- Look for **clusters** of close-together types

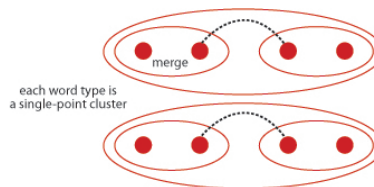
Plot in k dimensions ($k=3$)



Bottom-Up Clustering

- Start with one cluster per point
- Repeatedly merge 2 closest clusters
 - **Single-link:** $\text{dist}(A,B) = \min \text{dist}(a,b)$ for $a \in A, b \in B$
 - **Complete-link:** $\text{dist}(A,B) = \max \text{dist}(a,b)$ for $a \in A, b \in B$

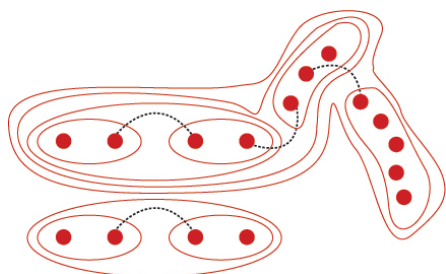
Single-Link



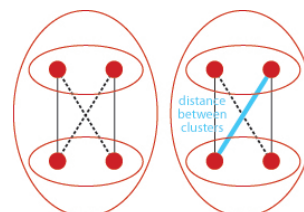
- Again, merge closest pair of clusters:
 - **Single-link:** clusters are close if any of their points are $\text{dist}(A,B) = \min \text{dist}(a,b)$ for $a \in A, b \in B$

Single-Link

- Fast, but tend to get long, stringy, meandering clusters



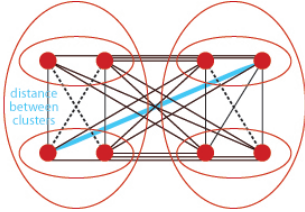
Complete-Link



- Again, merge closest pair of clusters:
 - **Complete-link:** clusters are close only if **all** of their points are $\text{dist}(A,B) = \max \text{dist}(a,b)$ for $a \in A, b \in B$

Complete-Link

- Slow to find closest pair – need quadratically many distances



Summary

- Start with one cluster per point
- Repeatedly merge 2 closest clusters
 - **Single-link:** $\text{dist}(A,B) = \min \text{dist}(a,b)$ for $a \in A, b \in B$
 - **Complete-link:** $\text{dist}(A,B) = \max \text{dist}(a,b)$ for $a \in A, b \in B$
 - too slow to update cluster distances after each merge; but alternatives!
 - **Average-link:** $\text{dist}(A,B) = \text{mean dist}(a,b)$ for $a \in A, b \in B$
 - **Centroid-link:** $\text{dist}(A,B) = \text{dist}(\text{mean}(A), \text{mean}(B))$
- Stop when clusters are “big enough”
 - e.g., provide adequate support for backoff (on a development corpus)
- Some flexibility in defining $\text{dist}(a,b)$
 - Might not be Euclidean distance; e.g., use vector angle

EM Clustering (for k clusters)

- **EM algorithm**
 - **Viterbi version** – called “k-means clustering”
 - **Full EM version** – called “Gaussian mixtures”
- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters
- **Parameters:** k points representing cluster centers
- **Hidden structure:** for each data point (word type), which center generated it?