Quiz #2

Topics
- Linguistics 101
- Parsing
  - Grammars, CFGs, PCFGs
  - Top-down vs. bottom-up
  - CKY algorithm
  - Grammar learning
  - Evaluation
  - Improved models
- Text similarity
  - Will also be covered on Quiz #3, though

Text Similarity

A common question in NLP is how similar are texts

\[
sim(\text{ }, \text{ }) = ?
\]

score:

rank: ?
Bag of words representation

For now, let's ignore word order:

Obama said banana repeatedly last week on tv, “banana, banana, banana”

(4, 1, 1, 0, 0, 1, 0, 0, …)

"Bag of words representation": multi-dimensional vector, one dimension per word in our vocabulary

Vector based word

Multi-dimensional vectors, one dimension per word in our vocabulary

Frequency of word occurrence

TF-IDF

One of the most common weighting schemes

TF = term frequency

IDF = inverse document frequency

\[ a' = a_i \times \log N / df_i \]

We can then use this with any of our similarity measures!

Normalized distance measures

Cosine

\[ sim_{cos}(A,B) = A \cdot B = \sum_{i=1}^{n} a_i b_i / \left( \sqrt{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2} \right) \]

L2

\[ dist_{L2}(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \]

L1

\[ dist_{L1}(A,B) = \sum_{i=1}^{n} |a_i - b_i| \]

\( a' \) and \( b' \) are length normalized versions of the vectors
Our problems

Which of these have we addressed?

- word order
- length
- synonym
- spelling mistakes
- word importance
- word frequency

A model of word similarity!

Word overlap problems

A: When the defendant and his lawyer walked into the court, some of the victim supporters turned their backs to him.

B: When the defendant walked into the courthouse with his attorney, the crowd turned their backs on him.

Word similarity

How similar are two words?

\[ \text{score: } \text{sim}(w_1, w_2) = ? \]

<table>
<thead>
<tr>
<th>rank</th>
<th>w</th>
<th>1</th>
<th>w_2</th>
<th>w_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>list</td>
<td>w_1 and w_2 are synonyms</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Word similarity applications

- General text similarity
- Thesaurus generation
- Automatic evaluation
- Text-to-text
  - paraphrasing
  - summarization
  - machine translation
- information retrieval (search)
Word similarity

How similar are two words?

score: \( \text{sim}(w_1, w_2) = ? \)

rank: \( w \quad ? \quad w_2 \quad w_3 \)

list: \( w_1 \) and \( w_2 \) are synonyms

Character-based similarity

\( \text{sim}(\text{turned}, \text{truned}) = ? \)

How might we do this using only the words (i.e. no outside resources?)

Word similarity

Four categories of approaches (maybe more)

- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

Edit distance (Levenshtein distance)

The edit distance between \( w_1 \) and \( w_2 \) is the minimum number of operations to transform \( w_1 \) into \( w_2 \)

Operations:
- insertion
- deletion
- substitution

\( \text{EDIT}(\text{turned}, \text{truned}) = ? \)
\( \text{EDIT}(\text{computer}, \text{commuter}) = ? \)
\( \text{EDIT}(\text{banana}, \text{apple}) = ? \)
\( \text{EDIT}(\text{wombat}, \text{worcester}) = ? \)
Edit distance

- Edit distance ($\text{EDIT}(\text{turned, truned}) = 2$
  - delete $u$
  - insert $u$

- Edit distance ($\text{EDIT}($computer, commuter$) = 1$
  - replace $p$ with $m$

- Edit distance ($\text{EDIT}($banana, apple$) = 5$
  - delete $b$
  - replace $n$ with $p$
  - replace $a$ with $p$
  - replace $n$ with $l$
  - replace $a$ with $e$

- Edit distance ($\text{EDIT}($wombat, worcester$) = 6$

Better edit distance

- Are all operations equally likely?
  - No

  Improvement: give different weights to different operations
  - replacing $a$ for $e$ is more likely than $z$ for $y$

- Ideas for weightings?
  - Learn from actual data (known typos, known similar words)
  - Intuitions: phonetics
  - Intuitions: keyboard configuration

Vector character-based word similarity

Similarity ($\text{sim}(\text{turned, truned}) = ?$

Any way to leverage our vector-based similarity approaches from last time?

Vector character-based word similarity

Similarity ($\text{sim}(\text{turned, truned}) = ?$

Generate a feature vector based on the characters (or could also use the set based measures at the character level)

Problems?
Vector character-based word similarity

\[
sim(\text{restful, fluster}) = ?
\]

<table>
<thead>
<tr>
<th>aa: 0</th>
<th>ae: 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab: 0</td>
<td>af: 0</td>
</tr>
<tr>
<td>ac: 0</td>
<td>ag: 0</td>
</tr>
<tr>
<td>ae: 1</td>
<td>ah: 1</td>
</tr>
<tr>
<td>af: 1</td>
<td>al: 1</td>
</tr>
<tr>
<td>ag: 0</td>
<td>an: 0</td>
</tr>
</tbody>
</table>

Character level loses a lot of information

Use character bigrams or even trigrams

WordNet

Lexical database for English
- 155,287 words
- 206,941 word senses
- 1,177,659 synsets (synonym sets)
- ~400K relations between senses
- Parts of speech rows, verbs, adjectives, adverbs

Word graph, with word senses as nodes and edges as relationships

Psycholinguistics
- W2V attempts to model human lexical memory
- Design based on psycholinguistic testing

Created by researchers at Princeton
- http://wordnet.princeton.edu/
- Lots of programmatic interfaces

Four general categories
- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts
WordNet relations

- synonym
- antonym
- hypernyms
- hyponyms
- holonym
- meronym
- troponym
- entailment
- (and a few others)

WordNet relations

- synonym — X and Y have similar meaning
- antonym — X and Y have opposite meanings
- hypernyms — subclass
  - beagle is a hypernym of dog
- hyponyms — superclass
  - dog is a hyponym of beagle
- holonym — contains part
  - car is a holonym of wheel
- meronym — part of
  - wheel is a meronym of car

WordNet relations

troponym — for verbs, a more specific way of doing an action
- run is a troponym of move
- dice is a troponym of cut

etailment — for verbs, one activity leads to the next
- sleep is entailed by snore

(And a few others)

WordNet

Graph, where nodes are words and edges are relationships
There is some hierarchical information, for example with hyp-er/o-nomy
To utilize WordNet, we often want to think about some graph-based measure.

Rank the following based on similarity:
- SIM(wolf, dog)
- SIM(wolf, amphipthon)
- SIM(wolf, stallion)
- SIM(dachshund, terrier)
What information/heuristics did you use to rank these?

- path length is important (but not the only thing)
- words that share the same ancestor are related
- words lower down in the hierarchy are finer grained and therefore closer

Utilizing information content:
- information content of the lowest common parent (Resnik, 1995)
- information content of the words minus information content of the lowest common parent (Jiang and Conrath, 1997)
- information content of the lowest common parent divided by the information content of the words (Lin, 1998)
Word similarity

Four general categories
- Character-based
  - turned vs. truned
cognates (night, nacht, nicht, natt, not, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

Dictionary-based similarity

Utilize our text similarity measures

\[
sim(\text{dog, beagle}) = \sim(\text{ ), beagle, dog})
\]

Example:
- Word similarity
  - \text{ ventured vs. traverse, night, nacht, nicht, natt, not, noc, noch, Orycteropus afer, feeding on ants and termites and having a long, extensile tongue, strong claws, and long ears, aardvark, One of a breed of small hounds having long ears, short legs, and a usually black, tan, and white coat, beagle, Any carnivore of the family Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.)

What about words that have multiple senses/parts of speech?
**Dictionary-based similarity**

1. part of speech tagging
2. word sense disambiguation
3. most frequent sense
4. average similarity between all senses
5. max similarity between all senses
6. sum of similarity between all senses

**Dictionary + WordNet**

WordNet also includes a “gloss” similar to a dictionary definition.

Other variants include the overlap of the word senses as well as those word senses that are related (e.g. hypernym, hyponym, etc.)
- incorporates some of the path information as well
- Banerjee and Pedersen, 2003

**Word similarity**

Four general categories
- Character-based
  - turned vs. trunad
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

**Corpus-based approaches**

<table>
<thead>
<tr>
<th>Word</th>
<th>ANY blurb with the word</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td></td>
</tr>
<tr>
<td>beagle</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
</tr>
</tbody>
</table>
The Beagle is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter legs.

Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern Beagle can be traced in Ancient Greece back to around the 5th century BC.

From medieval times, beagle was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard Beagle type was beginning to develop: the distinction between the North Country Beagle and Southern Beagle.

Word-context co-occurrence vectors

The Beagle is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter legs.

Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern Beagle can be traced in Ancient Greece back to around the 5th century BC.

From medieval times, beagle was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard Beagle type was beginning to develop: the distinction between the North Country Beagle and Southern Beagle.

Corpus-based: feature extraction

We'd like to utilize our vector-based approach.

How could we create a vector from these occurrences?

- Collect word counts from all documents with the word in it.
- Collect word counts from all sentences with the word in it.
- Collect all word counts from all words within X words of the word
- Collect all word counts from words in specific relationships: subject-object, etc.

Often do some preprocessing like lowercasing and removing stop words.
Corpus-based similarity

\[ \text{sim}(\text{dog, beagle}) = \text{sim}(\text{context}_\text{vector}(\text{dog}), \text{context}_\text{vector}(\text{beagle})) \]

The:
- the: 5
- is: 1
- as: 4
- breed: 2
- are: 1
- intelligent: 5

Web-based similarity

Web-based similarity

Concatenate the snippets for the top \( N \) results

Concatenate the web page text for the top \( N \) results
Another feature weighting

TF-IDF weighting takes into account the general importance of a feature

For distributional similarity, we have the feature \( f \), but we also have the word itself \( w \) that we can use for information

\[
\text{sim}(\text{context_vector}(\text{dog}), \text{context_vector}(\text{beagle}))
\]

<table>
<thead>
<tr>
<th>the.</th>
<th>2</th>
<th>the.</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>is.</td>
<td>1</td>
<td>is.</td>
<td>1</td>
</tr>
<tr>
<td>breeds.</td>
<td>2</td>
<td>breed.</td>
<td>1</td>
</tr>
<tr>
<td>area.</td>
<td>1</td>
<td>area.</td>
<td>1</td>
</tr>
<tr>
<td>are.</td>
<td>1</td>
<td>are.</td>
<td>1</td>
</tr>
<tr>
<td>intelligent</td>
<td>1</td>
<td>intelligent</td>
<td>1</td>
</tr>
<tr>
<td>and.</td>
<td>1</td>
<td>and.</td>
<td>1</td>
</tr>
<tr>
<td>to.</td>
<td>1</td>
<td>to.</td>
<td>1</td>
</tr>
<tr>
<td>modern</td>
<td>1</td>
<td>modern</td>
<td>1</td>
</tr>
</tbody>
</table>

Another feature weighting

Feature weighting ideas given this additional information?

\[
\text{sim}(\text{context_vector}(\text{dog}), \text{context_vector}(\text{beagle}))
\]

<table>
<thead>
<tr>
<th>the.</th>
<th>2</th>
<th>the.</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>is.</td>
<td>1</td>
<td>is.</td>
<td>1</td>
</tr>
<tr>
<td>breeds.</td>
<td>2</td>
<td>breed.</td>
<td>1</td>
</tr>
<tr>
<td>area.</td>
<td>1</td>
<td>area.</td>
<td>1</td>
</tr>
<tr>
<td>are.</td>
<td>1</td>
<td>are.</td>
<td>1</td>
</tr>
<tr>
<td>intelligent</td>
<td>1</td>
<td>intelligent</td>
<td>1</td>
</tr>
<tr>
<td>and.</td>
<td>1</td>
<td>and.</td>
<td>1</td>
</tr>
<tr>
<td>to.</td>
<td>1</td>
<td>to.</td>
<td>1</td>
</tr>
<tr>
<td>modern</td>
<td>1</td>
<td>modern</td>
<td>1</td>
</tr>
</tbody>
</table>

Another feature weighting

Mutual information

A bit more probability 😊

\[
I(X,Y) = \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\]

When will this be high and when will this be low?

\[
\text{sim}(\text{context_vector}(\text{dog}), \text{context_vector}(\text{beagle}))
\]

count how likely feature \( f \), and word \( w \) are to occur together
- incorporates co-occurrence
- but also incorporates how often \( w \) and \( f \), occur in other instances

Does IDF capture this?

Not really. IDF only accounts for \( f \), regardless of \( w \)
Mutual information

A bit more probability 😊

\[ I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]

if \( x \) and \( y \) are independent (i.e. one occurring doesn't impact the other occurring) then:

\[ p(x,y) = p(x)p(y) \]

What does this do to the sum?

Mutual information

A bit more probability 😊

\[ I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]

if \( x \) and \( y \) are independent (i.e. one occurring doesn't impact the other occurring) then:

\[ p(x,y) = p(x)p(y) \]

What is this asking?
When is this high?

How much more likely are we to see \( y \) given \( x \) has a particular value!
Point-wise mutual information

Mutual information

\[ I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]

Point-wise mutual information

\[ PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \]

How related are two variables (i.e. over all possible values/events)

PMI weighting

Mutual information is often used for feature selection in many problem areas

PMI weighting weights co-occurrences based on their correlation (i.e. high PMI)

How do we calculate these?

context_vector(beagle)

- the: 2
- is: 1
- a: 2
- breed: 1
- are: 1
- intelligent: 1
- and: 1
- to: 1
- modern: 1
- ...