Text Clustering

1

1

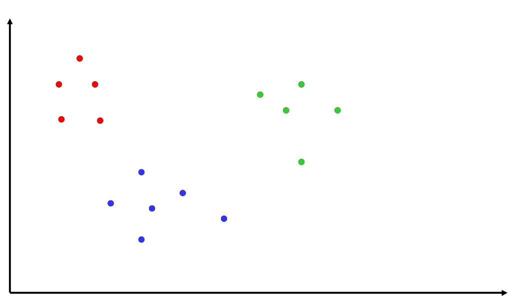
Clustering

- Partition unlabeled examples into disjoint subsets of *clusters*, such that:
  - Examples within a cluster are very similar
  - Examples in different clusters are very different
- Discover new categories in an *unsupervised* manner (no sample category labels provided).

2

2

Clustering Example

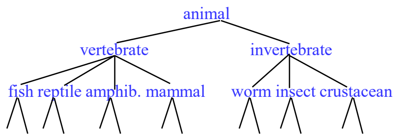


3

3

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of unlabeled examples.



- Recursive application of a standard clustering algorithm can produce a hierarchical clustering.

4

4

Aglommerative vs. Divisive Clustering

- *Aglommerative* (*bottom-up*) methods start with each example in its own cluster and iteratively combine them to form larger and larger clusters.
- *Divisive* (*partitional, top-down*) separate all examples immediately into clusters.

5

5

Hierarchical Agglomerative Clustering (HAC)

- Assumes a *similarity function* for determining the similarity of two instances.
- Starts with all instances in a separate cluster and then repeatedly joins the two clusters that are most similar until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

6

6

HAC Algorithm

Start with all instances in their own cluster.
Until there is only one cluster:
Among the current clusters, determine the two clusters, c_i and c_j, that are most similar.
Replace c_i and c_j with a single cluster $c_i \cup c_j$

7

7

Cluster Similarity

- Assume a similarity function that determines the similarity of two instances: *sim*(x,y).
  - Cosine similarity of document vectors.
- How to compute similarity of two clusters each possibly containing multiple instances?
  - Single Link: Similarity of two most similar members.
  - Complete Link: Similarity of two least similar members.
  - Group Average: Average similarity between members.

8

8

Non-Hierarchical Clustering

- Typically must provide the number of desired clusters, k.
- Randomly choose k instances as *seeds*, one per cluster.
- Form initial clusters based on these seeds.
- Iterate, repeatedly reallocating instances to different clusters to improve the overall clustering.
- Stop when clustering converges or after a fixed number of iterations.

9

9

K-Means

- Assumes instances are real-valued vectors.
- Clusters based on *centroids*, *center of gravity*, or mean of points in a cluster, *c*:

$$\vec{\mu}(\mathrm{c}) = \frac{1}{|c|}\sum_{\vec{x} \in c} \vec{x}$$

- Reassignment of instances to clusters is based on distance to the current cluster centroids.

10

10

Distance Metrics

- Euclidian distance (L_2 norm):

$$L_2(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

- L_1 norm:

$$L_1(\vec{x}, \vec{y}) = \sum_{i=1}^{m} |x_i - y_i|$$

- Cosine Similarity (transform to a distance by subtracting from 1):

$$1 - \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

11

11

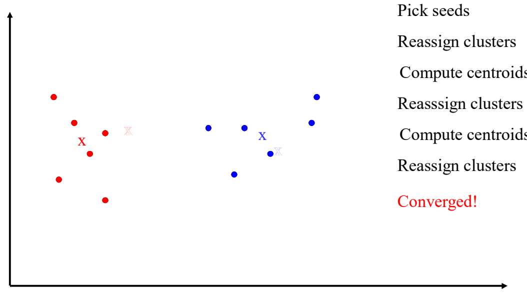
K-Means Algorithm

Let *d* be the distance measure between instances.
Select *k* random instances $\{s_1, s_2, \ldots s_k\}$ as seeds.
Until clustering converges or other stopping criterion:
  For each instance x_i:
    Assign x_i to the cluster c_j such that $d(x_i, s_j)$ is minimal.
  (Update the seeds to the centroid of each cluster)
  For each cluster c_j
    $s_j = \mu(c_j)$

12

12

K Means Example
(K=2)



Pick seeds
Reassign clusters
Compute centroids
Reasssign clusters
Compute centroids
Reassign clusters
Converged!

13

13

Information Extraction

14

14

Information Extraction (IE)

- Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
  - Newspaper articles
  - Web pages
  - Scientific articles
  - Newsgroup messages
  - Classified ads
  - Medical notes

15

15

MUC

- DARPA funded significant efforts in IE in the early to mid 1990's.
- Message Understanding Conference (MUC) was an annual event/competition where results were presented.
- Focused on extracting information from news articles:
  - Terrorist events
  - Industrial joint ventures
  - Company management changes
- Information extraction of particular interest to the intelligence community (CIA, NSA).

16

16

Other Applications

- Job postings
- Job resumes
- Seminar announcements
- Company information from the web
- Apartment rental ads
- Molecular biology information from MEDLINE

17

17

Sample Job Posting

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

18

18

Extracted Job Template

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

19

19

Amazon Book Description

```
….
</td></tr>
</table>
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>
<font face=verdana,arial,helvetica size=-1>
by <a href="/exec/obidos/search-handle-url/index=books&field-author=
            Kurzweil%2C%20Ray/002-6235079-4593641">
Ray Kurzweil</a><br>
</font>
<br>
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">
<img src="http://images.amazon.com/images/P/0140282025.01.MZZZZZZZ.gif" width=90
   height=140 align=left border=0></a>
<font face=verdana,arial,helvetica size=-1>
<span class="small">
<span class="small">
<b>List Price:</b> <span class=listprice>$14.95</span><br>
<b>Our Price: <font color=#990000>$11.96</font></b><br>
<b>You Save:</b> <font color=#990000><b>$2.99 </b>
(20%)</font><br>
</span>
<p> <br>…
```

20

20

Extracted Book Template

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence
Author: Ray Kurzweil
List-Price: $14.95
Price: $11.96
:
:

21

21

Web Extraction

- Many web pages are generated automatically from an underlying database.
- Therefore, the HTML structure of pages is fairly specific and regular (*semi-structured*).
- However, output is intended for human consumption, not machine interpretation.
- An IE system for such generated pages allows the web site to be viewed as a structured database.
- An extractor for a semi-structured web site is sometimes referred to as a *wrapper*.
- Process of extracting from such pages is sometimes referred to as *screen scraping*.

22

22

Learning for IE

- Writing accurate patterns for each slot for each domain (e.g. each web site) requires laborious software engineering.
- Alternative is to use machine learning:
  - Build a training set of documents paired with human-produced filled extraction templates.
  - Learn extraction patterns or a neural network to identify the fillers of each slot using an appropriate machine learning algorithm.

23

23

Evaluating IE Accuracy

- Always evaluate performance on independent, manually-annotated test data not used during system development.
- Measure for each test document:
  - Total number of correct extractions in the solution template: N
  - Total number of slot/value pairs extracted by the system: E
  - Number of extracted slot/value pairs that are correct (i.e. in the solution template): C
- Compute average value of metrics adapted from IR:
  - Recall = C/N
  - Precision = C/E
  - F-Measure = Harmonic mean of recall and precision

24

24

Semantic Parsing
for
Question Answering

25

25

Semantic Parsing

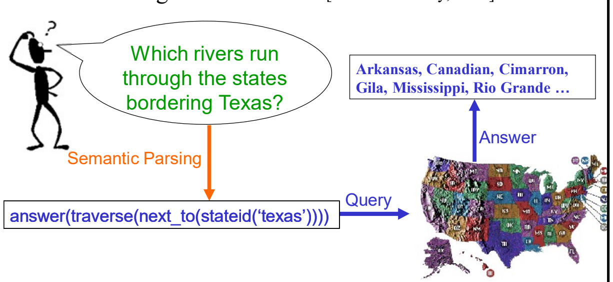
- ***Semantic Parsing***: Transforming natural language (NL) sentences into completely formal ***logical forms*** or ***meaning representations*** (MRs).
- Sample application domains where MRs are directly executable by another computer system to perform some task.
  - Database/knowledge-graph queries
  - Robot command language

26

26

Geoquery: A Database Query Application

- Query application for U.S. geography database containing about 800 facts [Zelle & Mooney, 1996]



27

27

Formal Query Language

- Most early work on computational semantics is based on predicate logic

What is the smallest state by area?
answer(x_1,smallest(x_2,(state(x_1),area(x_1,x_2))))

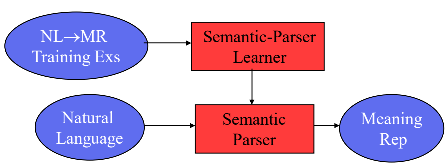
x_1 is a logical variable that denotes "the smallest state by area"

- More recent work uses deep neural nets to directly map "language to code" and generate SQL queries or other programs

28

Learning Semantic Parsers

- Manually programming robust semantic parsers is difficult due to the complexity of the task.
- Semantic parsers can be learned automatically from sentences paired with their logical form.



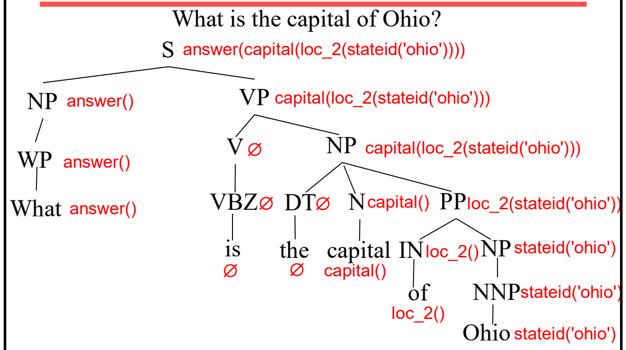
29

Compositional Semantics

- Approach to semantic analysis based on building up an MR compositionally based on the syntactic structure of a sentence.
- Build MR recursively bottom-up from the parse tree.

```
BuildMR(parse-tree)
    If parse-tree is a terminal node (word) then
        return an atomic lexical meaning for the word.
    Else
      For each child, subtree_i, of parse-tree
          Create its MR by calling BuildMR(subtree_i)
    Return an MR by properly combining the resulting MRs
      for its children into an MR for the overall parse-tree.
```

Composing MRs from Parse Trees



31

31

Experimental Corpora

- GeoQuery [Zelle & Mooney, 1996]
  - 250 queries for the given U.S. geography database
  - 6.87 words on average in NL sentences
  - 5.32 tokens on average in formal expressions
  - Also translated into Spanish, Turkish, & Japanese.

32

32

Experimental Methodology

- Evaluated using standard 10-fold cross validation
- Correctness
  - CLang: output ***exactly matches*** the correct representation
  - Geoquery: the resulting query retrieves the same answer as the correct representation
- Metrics

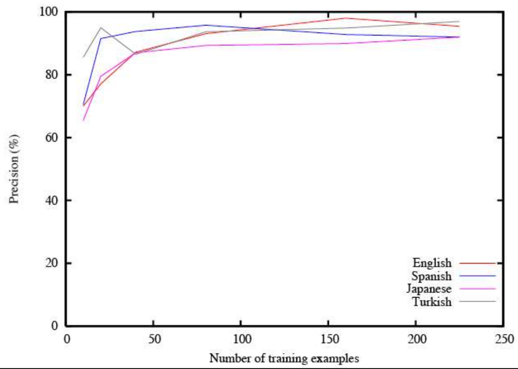
$$Precision = \frac{|Correct\ Completed\ Parses|}{|Completed\ Parses|}$$

$$Recall = \frac{|Correct\ Completed\ Parses|}{|Sentences|}$$

33

33

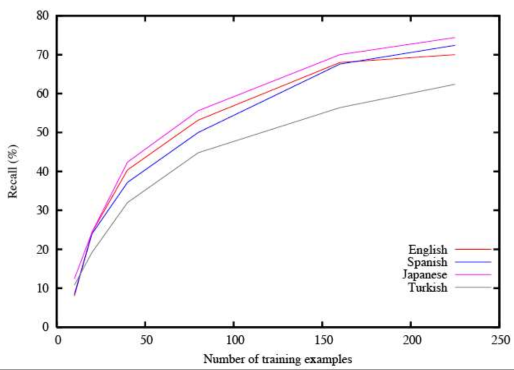
Precision Learning Curve for GeoQuery (WASP)



34

34

Recall Learning Curve for GeoQuery (WASP)



35

35