Data Mining Text - Natural Language Processing (NLP)

-Alan Turing 1950 article in British journal Mind – "Computing Machinery and Intelligence"
Posed what is now known at the "Turing test"
Can machines think?
Replaces that question with the "Imitation Game"
Computer interacts by written language with a judge
If judge cannot reliably determine that it's computer
Computer wins

-Development based on linguistic theory of grammars

-Until 1980's work mostly directed towards development of semantic rules for tasks like automatic translation and computer "understanding".

-"Likelihood of a sentence" had no meaning.

Machine Learning Applied to NLP

-Starting in 1980's work turned toward machine learning approaches

-Probabilistic decisions replaced hard rules.

-Early successes in machine translation

-Able to work with large existing document corpora (versus requiring purpose-built documents)

-More robust on previously unseen examples

-More robust in the face of errors

Example – Part of Speech (POS) Tagger

-Task is to determine parts of speech (noun, verb, adjective, conjunction, etc.) for words in a sentence

-Two parts to the task – Training and Evaluation (or deployment)

1. Training – Given corpus of marked sentences develop probabilistic model for word POS, given the possibilities for the word and the POS of the words surrounding it. Make the model complex enough to perform well on training data and simple enough that the performance generalizes to new data.

2. Evaluate on unseen data and deploy

Advantages of ML Approach

1. ML approach focuses on frequently occurring patterns versus tendency for hand written rules to focus on pathological "corner" cases.

2. Statistical approach more robust to unseen examples, unfamiliar structures and errors in input (which are frequent).

3. ML approaches scale. Algorithms map onto systems (Map Reduce) for handling web-scale data. Hand-written rules constrained by human and prone to errors as rule set grows.

Major NLP Tasks

1. Automatic Summarization: Produce a readable summary of text (e.g. articles in the financial section of newspaper)

2. Coreference resolution: Given a body of text determine which words (called "mentions") refer to the same objects. Anaphora resolution is a specific example – find antecedents for pronouns.

3. Discourse Analysis: Discover the nature of discourse relationships between sentences (e.g. elaboration, explanation, contrast). Determine psycho social relationships from speech (e.g. power relationship). Classify speech acts (e.g. yes-no question, content question, statement, etc.)

4. Machine Translation: Translate written text in one language into written text in another.

5. Named Entity Recognition (NER): Given a stream of text, determine which items map to proper names identify the type (e.g. place, name, organization). Capitalization isn't always enough ((Abraham) Lincoln never visited Lincoln (Nebraska))

6. Natural Language Generation: Convert information from database into readable language.

7. Part of Speech Tagging: In natural language text identify noun, verb, conjunction, ronoun, etc.

8. Question answering: Given natural language question, generate natural language answer.

Major NLP Tasks – Cont.

9. Relationship Extraction: Given natural language text determine relationship between named entities

10. Sentiment Analysis: Extract subjective information from set of documents (e.g. determine response to product release from social media)

Software Packages for NLP

-R
tm
openNLP
lda
lsa

-Python
NLTK