Automatic Text Understanding of Content and Text Quality

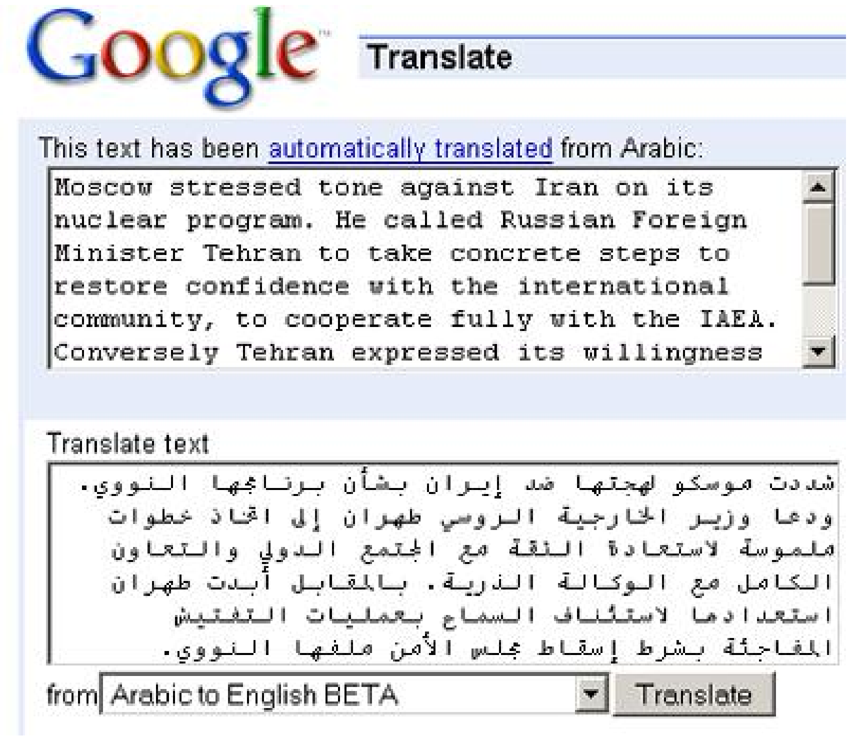
Ani Nenkova

University of Pennsylvania

Automatic semantics

- Computers help us navigate data
  - User query --- relevant documents
  - Object --- description

- Computers generate text
  - Machine translation
  - Automatic summarization



This text has been automatically translated from Arabic:

```
Moscow stressed tone against Iran on its
nuclear program. He called Russian Foreign
Minister Tehran to take concrete steps to
restore confidence with the international
community, to cooperate fully with the IAEA.
Conversely Tehran expressed its willingness
```

Translate text

شددت موسكو لهجتها ضد إيران بشأن برنامجها النووي.
ودعا وزير الخارجية الروسي طهران إلى اتخاذ خطوات
ملموسة لاستعادة الثقة مع المجتمع الدولي والتعاون
الكامل مع الوكالة الذرية. بالمقابل أبدت طهران
استعدادها لاستئناف السماح بعمليات التفتيش
المفاجئة بشرط إسقاط مجلس الأمن ملفها النووي.

from Arabic to English BETA

Two aspects of interpreting text

- What is the text about?
  - Meaning
  - Long tradition of research

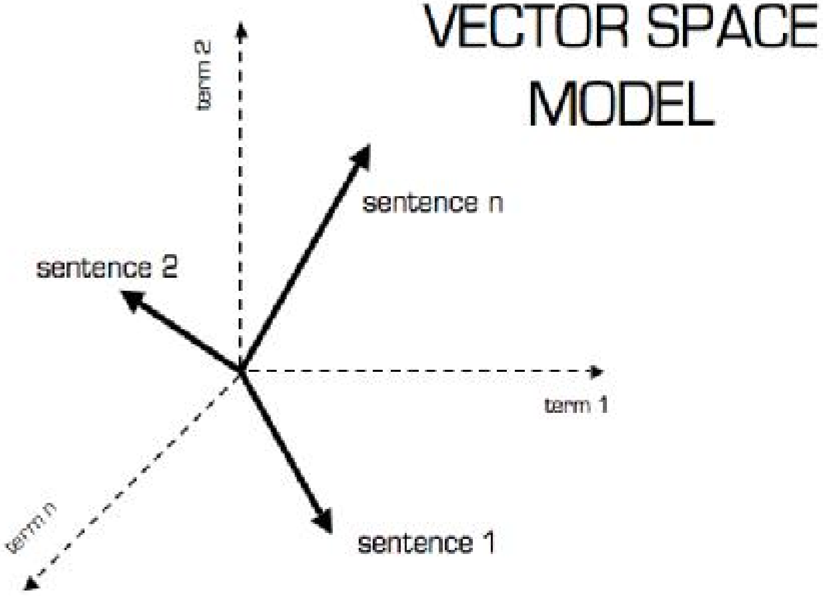
TOPIC

- How easy/pleasant is the text to read?
  - Text quality
  - Emerging as research area

QUALITY

Vector space model of semantics

- Represent objects as points in space
- Points near each other are semantically similar



Information retrieval

text represented as a vector of terms

D1 The key is in the front pocket of the backpack.

D2 The backpack is in the car .

$sim(q,d)$

	D1	D2
the	3	2
key	1	0
is	1	1
in	1	1
front	1	0
pocket	1	0
of	1	0
backpack	1	1

$$sim(D_1,D_2) = \frac{1}{dist(v_{D_1},v_{D_2})}$$

Cosine similarity

$$\cos(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

Word meaning

words as vectors of contexts

Context words that appear near the target word

Accumulated across many occurrences of the target word

Used for query expansion or lexicon construction

	red wine	green tea
drink	63	89
enjoy	48	20
beef	20	1
cookies	2	35
hot	3	100
california	89	2
dinner	79	40
morning	0	78

Words most similar to wine

beer, white wine, red wine, Chardonnay, champagne, fruit, food, coffee, juice, Cabernet…

Words most similar to excellent/amazing

cute, fabulous, top of the line, melt in the mouth

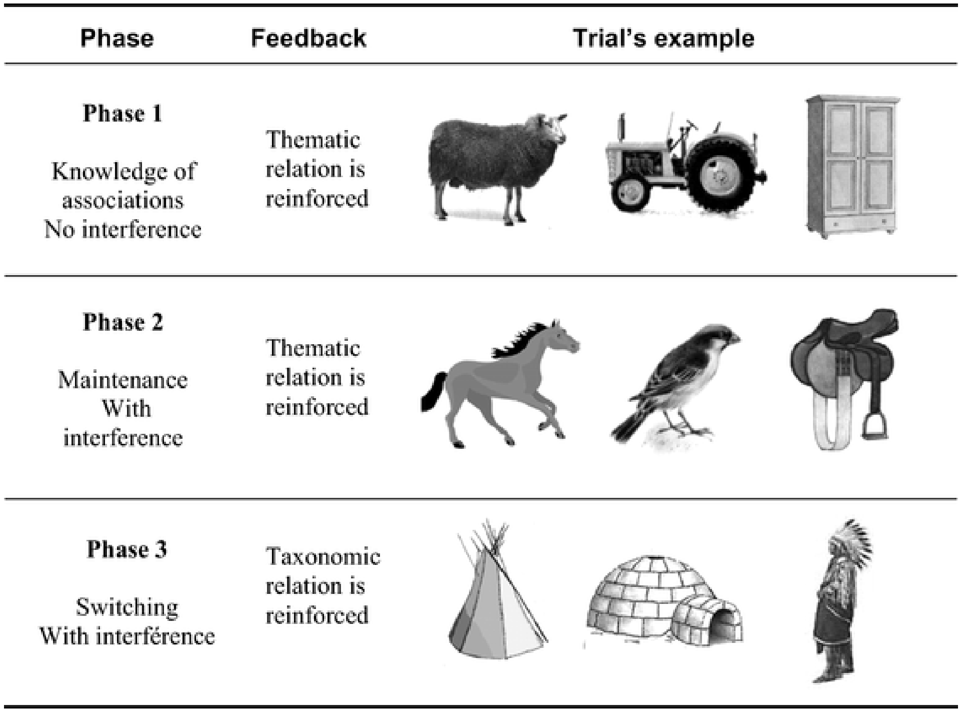
Words most similar to bad/horrible

subpar, crappy, out of touch, sick to my stomach

Relationships between words

pairs of words as vectors of patters

- Word pairs
  - mason:stone
  - carpenter:wood
- Patterns
  - ``X cuts Y”
  - ``X works with Y”
- Tasks
  - Finding similar patterns (paraphrases)
  - Finding words that share a semantic relationship



Phase	Feedback	Trial’s example
Phase 1 Knowledge of associations No interference	Thematic relation is reinforced	
Phase 2 Maintenance With interference	Thematic relation is reinforced	
Phase 3 Switching With interférence	Taxonomic relation is reinforced	

Some challenges

- Many documents and many words
  - More compact representations are necessary
- Reduce noise
  - Some entries in the representation are not meaningful
  - Find ways to eliminate them
- Richer contexts
  - Use concepts rather than words
  - Part of speech and syntactic information
- Choice of similarity metric
- Vector representations for machine learning

Text quality: three case studies

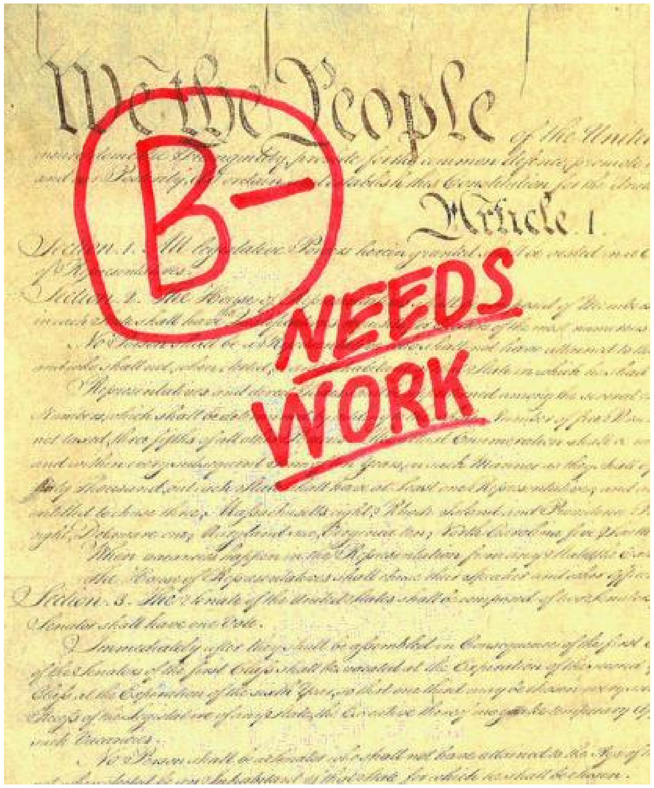
- Reproduce people’s ratings of quality
  - Identify correlates of well-written text

- Evaluate the linguistic quality of summaries
  - Currently done manually, expansive

- How specific or general is a sentence
  - A well-written text is a mix of both

How well-written is this text?

- Asked graduate students at Penn
- 30 Wall Street Journal articles



Not correlated with ratings

- Average characters per word
- Average number of words per sentence

- Subordinate clauses
  - Close to significant ``more is better”
- Similarity between adjacent sentences

Significant correlates of quality ratings

- Number of words (shorter is better)
- Text log likelihood
  - Word probabilities taken from newspaper text

$$\sum_{w} C(w) \log(P(w \mid News))$$

- Average number of verb phrases (more better)
- Discourse relations log likelihood

Discourse relations in text

COMPARISON

Mary likes to cook. Her husband prefers to eat out.

CONTINGENCY

He is a reliable person. I would choose him.

TEMPORAL

He decided not to go *after* he heard the forecast.

EXPANSION

The 40 year old Mr. Murakami is a publishing sensationin Japan.

A more recent novel, “Norwegian wood”, has sold more than forty million copies since Kodansha published it in 1987.

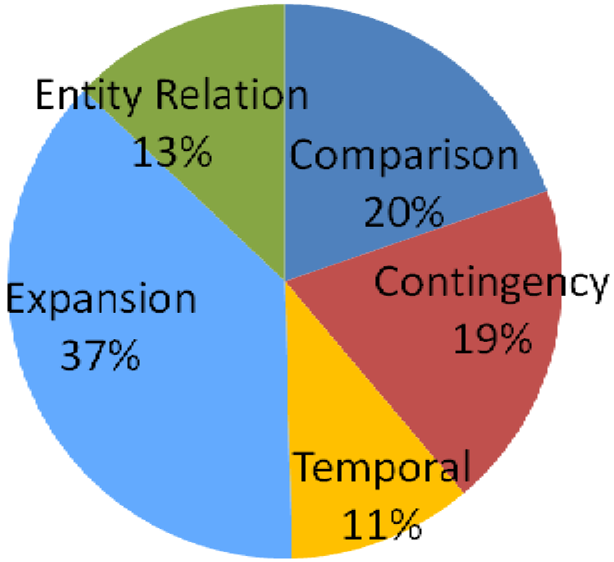
EXPLICIT

but, however, since, while, also connective

IMPLICIT

adjacent sentences, no

Distribution of implicit relations in Wall Street Journal articles



Features for automatic detection of discourse relations

- Co-occurrence of word pairs
- Sentiment words (positive, negative, neutral)
- Probabilities of words in the sentences
- Verb similarity
- Part of speech and syntactic information

Automatic evaluation of summary linguistic quality

- Annual evaluations conducted by NIST
  - 40—60 systems evaluated each year
  - On about 50 test inputs
- For content, automatic metrics rank systems similarly to the NIST evaluators
- Can linguistic quality be automatically rated?

Automatic quality rating accuracy

- Which of the summarizer is better over the entire test set?
  - 90% accurately predicted

- Which summary for a given text collection is better?
  - 65%--70% accurately predicted

Predicting sentence specificity

- A well-written text is a mix of general and specific sentences
- General sentences provide an overview and state topics
- Specific sentences provide details and support for the general claims

Example predictions

The novel, a story of a Scottish low-life narrated largely in Glaswegian dialect, is unlikely to prove a popular choice with booksellers who have damned all six books shortlisted for the prize as boring, elitist and – worse of all – unsaleable.

Specific

...

The Booker prize has, in its 26-year history, always provoked controversy.

General

Some of the features in the classifier

- Evaluative words
  - Occur more often in general text
- Syntax and part of speech
  - More adjectives and adverbs in general sentences
  - More plural nouns
- Numbers and proper names
- Word specificity
- Sentence length is not a useful feature

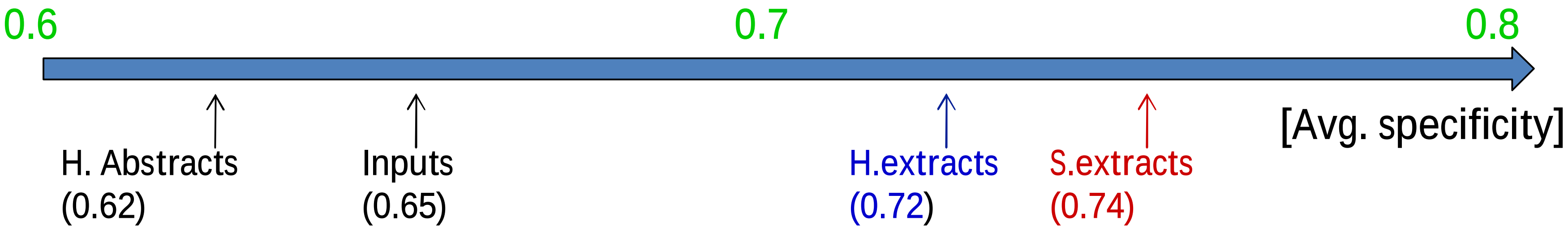
Accuracy of prediction

- People can make judgments about general/specific for the majority of sentences in text

- But there are some that are not as clear cut
  - So people do not agree with each other

- Classifier confidence correlates with human agreement
- Classifier accuracy is around 80% for any sentence
- 94% for sentences which people will agree on

Specificity analysis of summaries



0.6 0.7 0.8

[Avg. specificity]

H. Abstracts (0.62)

Inputs (0.65)

H.extracts (0.72)

S.extracts (0.74)

1. More general content is preferred in abstracts

2. Simply the process of extraction makes summaries more specific

3. System summaries are overly specific

Automatic summarization and specificity: pushing systems to understand more

- The more specific a summary is, the worse its content is judged
  - General sentences are like mini summaries

  Details of Maxwell's death were sketchy.

  Folksy was an understatement.

- The more general a summary is, the worst its linguistic quality is judged
  - Systems are not good at finding the proper context for general sentences

  With Tower's qualifications for the job, the nominations should have sailed through with flying colors. [Specific]

  Instead it sank like the Bismarck. [General]

Conclusions

- Many successful applications in content understanding

- More subtle understanding tasks have been drawing attention: discourse relations and sentence specificity

- Leading to progress in text quality research