Information Extraction: Overview

School of Information Studies
Syracuse University

Motivation for Information Extraction

When we covered semantic analysis, we focused on:

- The analysis of single sentences
- A deep approach that could, in principle, be used to extract considerable information from each sentence
- And a tight coupling with syntactic analysis

Unfortunately, when released in the wild, such approaches have difficulties with:

- **Speed:** deep syntactic and semantic analysis of each sentence is too slow for many applications
  - Transaction processing where large amounts of newly encountered text has to be analyzed
- **Coverage:** real-world texts tend to strain both the syntactic and semantic capabilities of most systems

Slides adapted from Jim Martin

Information Extraction

So, just as we can do with partial parsing and chunking for syntax, we can look for more lightweight techniques that get us most of what we might want in a more robust manner.

- Figure out the entities (the players, props, instruments, locations, etc. in a text).
- Figure out how they’re related.
- Figure out what they’re all up to.
- And do each of those tasks in a loosely coupled, data-driven manner.

Targeted Semantic Analysis

Ordinary newswire text is often used in typical examples.

- Target "who,” “what,” “when,” “where” of news events
- And there’s an argument that there are useful applications there

The real interest/money is in specialized domains.

- Bioinformatics
- Patent analysis
- Specific market segments for stock analysis
- Intelligence analysis
- Etc.

Overview of Information Extraction

CHICAGO (AP): Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Named Entity Recognition

Find the named entities and classify them by type.

CHICAGO (AP): Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Relation Extraction

Basic task: Find all the *classifiable* relations among the named entities in a text (populate a database).

- Employs (e.g., {<American, Tim Wagner> })
- Part-of (e.g., {<United, UAL>, {American, AMR} >)

CHICAGO (AP): Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Event Detection

Find and classify all the events of interest in a text.

- Most verbs introduce events/states, but not all (*give a kiss*)
- Nominalizations often introduce events
  - *Collision, destruction, the running…*

CHICAGO (AP): Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Temporal and Numerical Expressions

Find all the temporal expressions.

- Normalize them based on some reference point.

Find all the numerical expressions.

- Classify by type and normalize.

CHICAGO (AP): Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Template Analysis

Many news stories have a script-like flavor to them. They have fixed sets of expected events, entities, relations, etc.

Template, schema, or script processing involves:

- Recognizing that a story matches a known script
- Extracting the parts of that script

CHICAGO (AP): Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Information Extraction: Named Entity Recognition (NER)

School of Information Studies
Syracuse University

Information Extraction Typical Tasks

Named entity recognition and classification

Coreference analysis

Temporal and numerical expression analysis

Event detection and classification

Relation extraction

Template analysis

NER (Named Entity Recognition)

Find and **classify** all the named entities in a text.

What is a named entity?

- A mention of an entity using its name
    - *Kansas Jayhawks*
- This is a sub-set of the possible mentions
    - *Kansas, Jayhawks, the team, it, they*

Find means identify the exact span of the mention.

Classify means determine the category of the entity being referred to.

Typical Named Entity (NE) Types for Newswire Text

Some applications add more specific types.

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, automobiles

Examples of NE for Newswire Text

Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel.*
Vehicles	The updated *Mini Cooper* retains its charm and agility.

Ambiguity

A common ambiguity is whether a city, state, or country is a location or a political entity.

- *Palo Alto* could be either.

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

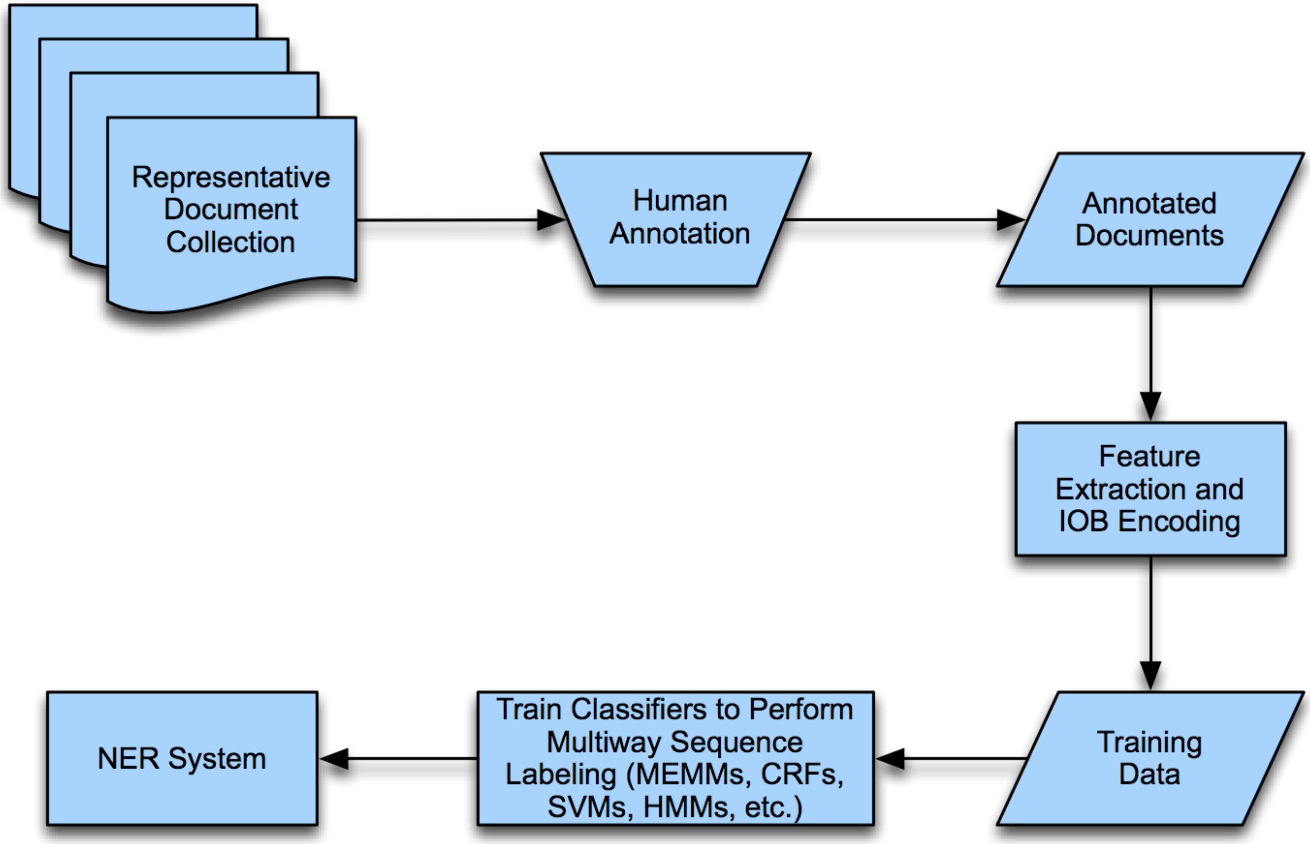
[$_{PERS}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up two games to one in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.
The [$_{FAC}$ *Washington*] had proved to be a leaky ship, every passage I made…

NER Approaches

As with partial parsing and chunking, there are two basic approaches (and hybrids).

- Rule-based (regular expressions)
  - Lists of names
  - Patterns to match things that look like names
  - Patterns to match the environments in which classes of names tend to occur
- ML-based approaches
  - Get annotated training data
  - Extract features
  - Train systems to replicate the annotation

ML Approach


Representative Document Collection
Human Annotation
Annotated Documents
Feature Extraction and IOB Encoding
Training Data
Train Classifiers to Perform Multiway Sequence Labeling (MEMMs, CRFs, SVMs, HMMs, etc.)
NER System

Encoding for Sequence Labeling

Named entity annotation often uses the IOB encoding.

- For *N* classes (i.e., types of entities), we have 2 * *N* + 1 tags for tokens.
  - B of that class for a token at the **beginning** of an entity of that class
  - I of that class for a token **inside** an entity of that class
  - O for a token **outside** of any class
- Each token in a text gets a tag.

NER Features

Features may include the word, POS tag, IOB for its phrase type, and the shape of the word.

- Feature examples

Features				Label
American	NNP	B_{NP}	cap	B_{0RG}
Airlines	NNPS	I_{NP}	cap	I_{0RG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{0RG}
Corp.	NNP	I_{NP}	cap-punc	I_{0RG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O

NER Features

Features may include the word, POS tag, IOB for its phrase type, and the shape of the word.

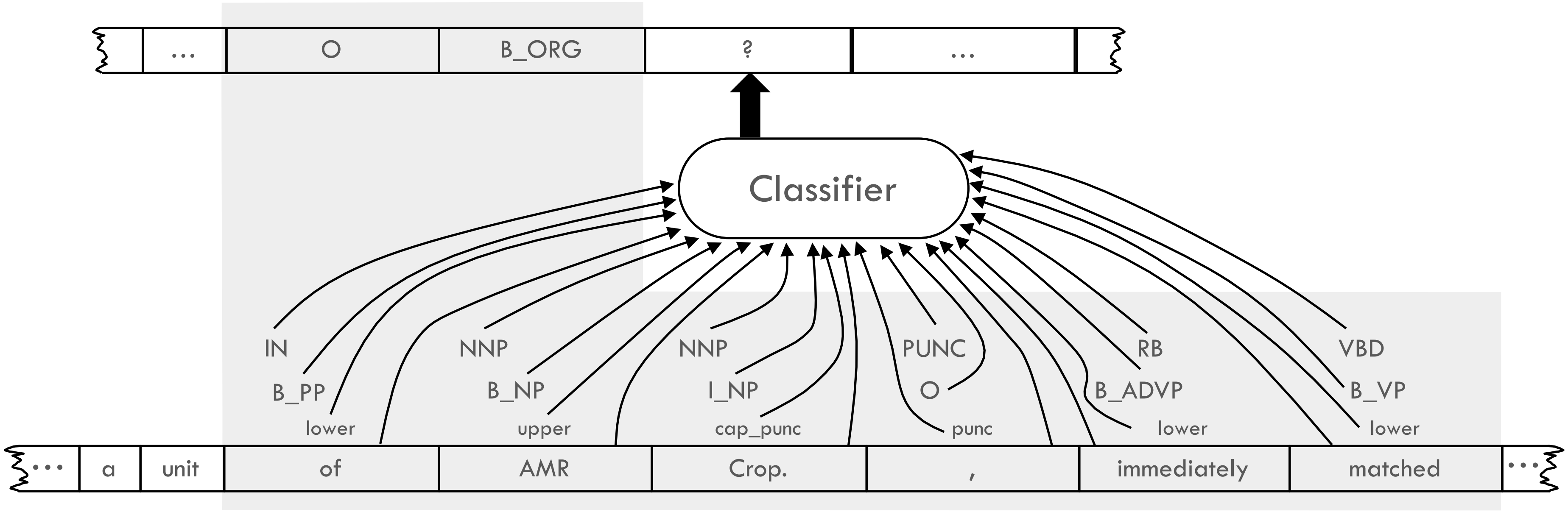
- More feature examples

Features				Label
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
Said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

NER as Sequence Labeling

A classifier may use words and features from tokens in the sequence before and after the one being classified.

- A sequential classifier may also use the predicted label of the previous token.



Information Extraction: Relations

School of Information Studies
Syracuse University

Relations

Once you have captured the entities in a text, you might want to ascertain how they relate to one another, according to the **relations of interest.**

- Here, we're just talking about explicitly stated relations.

Relation Extraction

CHICAGO (AP): Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

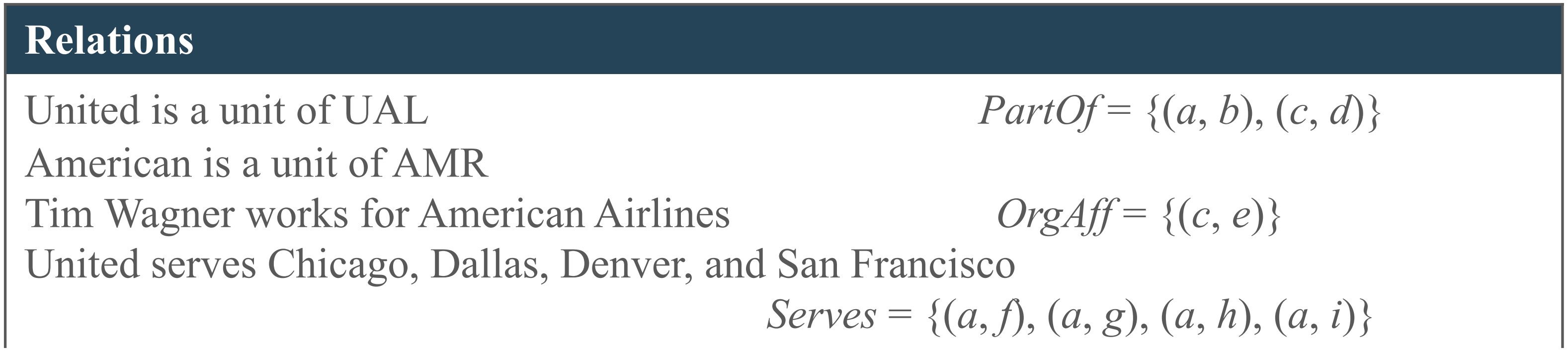
Relation Types

As with named entities, the list of relations is application-specific. For generic news texts...

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER → PER
	Organizational	*spokesman for, president of*	PER → ORG
	Artifactual	*owns, invented, produces*	(PER | ORG) → ART
Geo-spatial			
	Proximity	*near, on outskirts*	LOC → LOC
	Directional	*southeast of*	LOC → LOC
Part-of			
	Organizational	*a unit of, Parent of*	ORG → ORG
	Political	*annexed, acquired*	GPE → GPE

Relations

By relation, we really mean sets of tuples.



Relations

United is a unit of UAL $PartOf = \{(a, b), (c, d)\}$
American is a unit of AMR
Tim Wagner works for American Airlines $OrgAff = \{(c, e)\}$
United serves Chicago, Dallas, Denver, and San Francisco
$Serves = \{(a, f), (a, g), (a, h), (a, i)\}$

PartOf = {(United, UAL), (American, AMR)}
OrgAff = {(Tim Wagner, American Airlines)}
Serves = {(United, Chicago), (United, Dallas),
(United, Denver), (United, San Francisco)}

Relation Analysis

As with semantic role labeling, we can divide this task into two parts.

- Determining if two entities are related
- And if they are, classifying the relation

The reason for doing this is two-fold.

- Cutting down on training time for classification by eliminating most pairs
- Producing separate feature sets that are appropriate for each relation classification task

Relation Analysis

Let’s just worry about named entities within the same sentence.

- But, in a system, we will also use entities that are resolved by coreference to pronouns and other referring phrases.
- For every pair of entities in the sentence:

  Decide if the two entities are related (first classifier)
  Decide on the relation (second labeling classifier)

Features

We can group the features (for both tasks) into three categories:

1. **Features of the named entities involved**
   - Their types
     - Concatenation of the types
   - Head words of the entities
     - *George Washington Bridge*
   - Words in the entities

Features

2. **Features derived from the words between and around the named entities**
   - Particular positions to the left and right of the entities
     - +/– 1, 2, 3
     - Bag of words between
3. **Features derived from the syntactic environment that governs the two entities**
   - Constituent path through the tree from one to the other
   - Base syntactic chunk sequence from one to the other
   - Dependency path

Example

For the following example, we're interested in the possible relation between American Airlines and Tim Wagner.

- American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Entity-based features	
Entity$_1$ type	ORG
Entity$_1$head	*airlines*
Entity$_2$ type	PERS
Entity$_2$ head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ a, *unit*, of, *AMR*, *Inc*., *immediately*, *matched*, *the*, *spokesman* }
Word(s) before Entity$_1$	NONE
Word(s) after Entity$_2$	*said*
Syntactic features	
Constituent path	$NP \uparrow NP \uparrow S \downarrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Bootstrapping Approaches

What if you don’t have enough annotated text to train on?

- But you might have some seed tuples from the annotated text.
- Or you might have some patterns that work pretty well.

Can you use those seeds to do something useful?

- Co-training and active learning use the seeds to train classifiers to tag more data to train better classifiers.
- Bootstrapping tries to learn directly (populate a relation) through direct use of the seeds.

Bootstrapping Example: Seed Tuple

<Mark Twain, Elmira> seed tuple

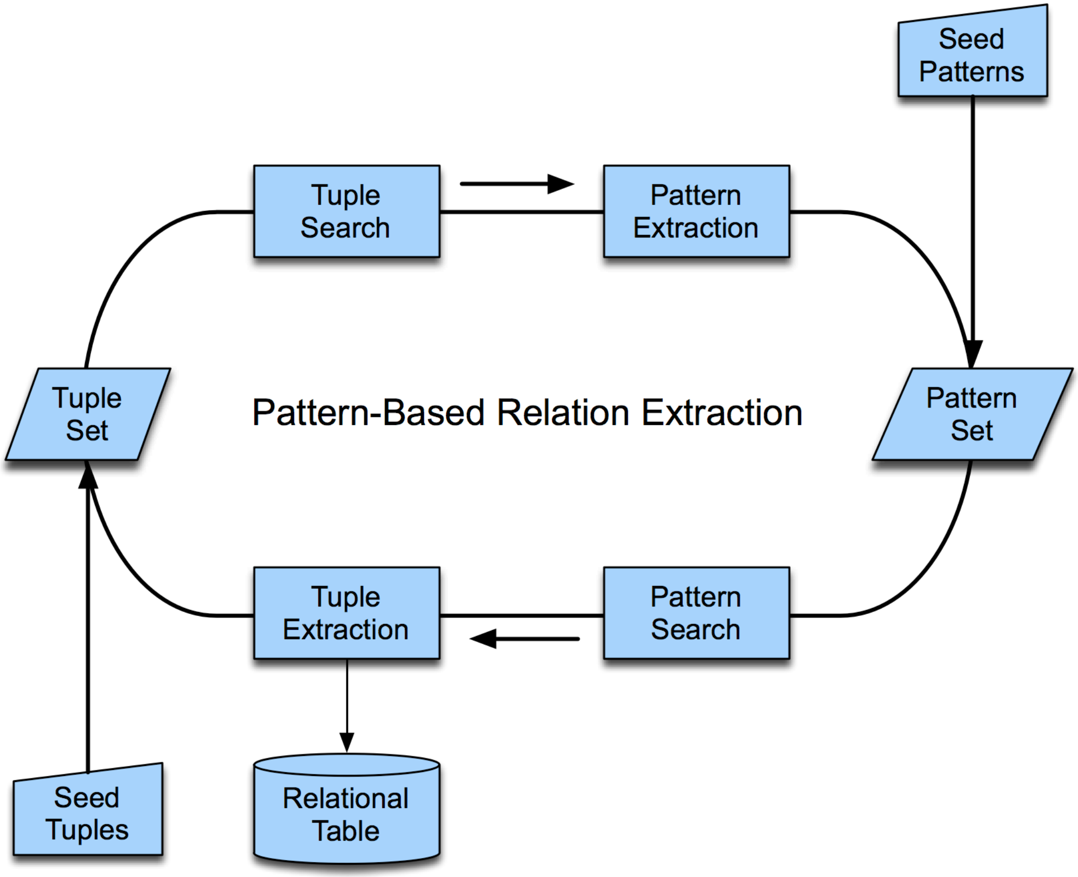
- For relation “Location of Burial”

- Search (Google)
- “Mark Twain is buried in Elmira, NY”
  - X is buried in Y
- “The grave of Mark Twain is in Elmira”
  - The grave of X is in Y
- “Elmira is Mark Twain’s final resting place”
  - Y is X’s final resting place

Use those patterns to Google for new tuples that you don’t already know

- Note that we can get some “noise” if there are other relations between Mark Twain and Elmira

Bootstrapping Relations



Template Filling

For stories/texts with stereotypical sequences of events, participants, props, etc.

Represent these facts as slots and slot-fillers: templates (frames, scripts, schemas)

- Evoke the right template
- Identify the story elements that fill each slot

Similar approaches as to relation extraction, except that you also have the option of developing patterns or classifiers for more than one slot at once

Airline Example

Example template for "Attempt to raise fares," with slots filled in from text

FARE-RAISE ATTEMPT:

LEAD AIRLINE :	UNITED AIRLINES
AMOUNT :	$6
EFFECTIVE DATE :	2006–10–26
FOLLOWER :	AMERICAN AIRLINES

Text Analysis Conference (TAC)

NIST is sponsoring these yearly text-analysis tasks (tracks) in the same spirit as TREC for information retrieval (IR)

Knowledge-based population (KBP)

- Also tracks on textual entailment and summarization

Participants must process news articles and prepare an information-extraction template formatted as a Wikipedia infobox

- Must also resolve entities across documents
- In 2010, must also detect "certainty" of information
- http://www.nist.gov/tac/

IE Example: Bioinformatic NLP

School of Information Studies
Syracuse University

Bioinformatic NLP

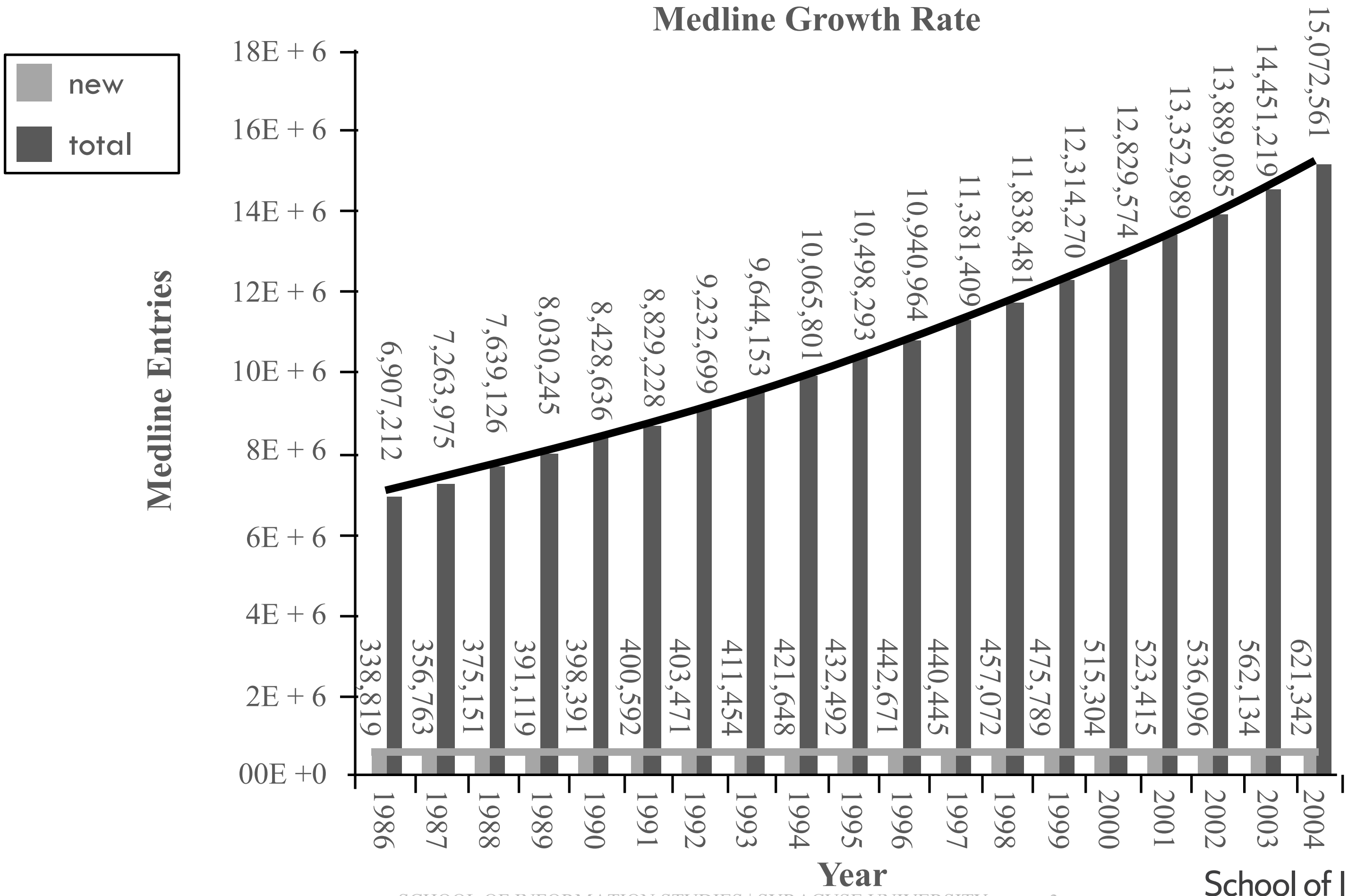
So far, we have been looking at newswire text with fairly standard names and relations.

- But names and relations in other domains may be very different

An example domain is bioinformatics.

- Very important
- Practitioners care about the technology
  - They have problems they're trying to solve
- A lot of text available
- A lot of interesting problems
- **Domain examples from Jim Martin**

A Lot of Text


Medline Growth Rate
new
total
Medline Entries
18E + 6
16E + 6
14E + 6
12E + 6
10E + 6
8E + 6
6E + 6
4E + 6
2E + 6
00E +0
6,907,212
7,263,975
7,639,126
8,030,245
8,428,636
8,829,228
9,232,699
9,644,153
10,065,801
10,498,293
10,940,964
11,381,409
11,838,481
12,314,270
12,829,574
13,352,989
13,889,085
14,451,219
15,072,561
338,819
356,763
375,151
391,119
398,391
400,592
403,471
411,454
421,648
432,492
442,671
440,445
457,072
475,789
515,304
523,415
536,096
562,134
621,342
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
Year

Problem Areas

Mainly variants of NER and relation analysis

- NER
  - Detecting and classifying named entities
  - And also **normalization**
    - Mapping that named entity to a particular entity in some external database or ontology
    - Medical field has external databases, or ontologies (e.g., UMLS for standard forms of named entities)
- Relation analysis
  - How various biological entities interact

Bio-NER

Large number of fairly specific types

Wide (really wide) variation in the naming of entities

- Gene names
  - *White, insulin, BRCA1, breast cancer associated 1, ether a go-go,* etc.

Semantic Class	Examples
Cell lines	*T98G, HeLa cell, Chinese hamster ovary cells, CHO cells*
Cell types	*primary T lymphocytes, natural killer cells, NK cells*
Chemicals	*citric acid, 1,2- diiodopentane, C*
Drugs	*cyclosporin A, CDDP*
Genes/proteins	*white, HSP60, protein kinase C, L23A*
Malignancies	*carcinoma, breast neoplasms*
Medical/clinical concepts	*amyotrophic lateralsclerosis*
Mouse strains	*LAFT, AKR*
Mutations	*C10T, Ala64* → *Gly*
Populations	*judo group*

Biorelations

Combination of IE- and SRL-style relation analysis

(22.27) [$_{THEME}$ Full-length cPLA2] was [$_{TARGET}$ phosphorylated] stoichiometrically by [$_{AGENT}$ p42 mitogen-activated protein (MAP) kinase] in vitro… and the major site of phosphorylation was identified by amino acid sequencing as [$_{SITE}$ Ser505]

Bioinformatic IE

Much work in NLP is concerned with portability and generality.

- How can we get systems trained on one genre/domain to work on a different one?

Biologists don't seem to care much about this.

- They're happy if you build a specific system to solve their specific problem.

Introduction to Machine Translation

School of Information Studies
Syracuse University

Machine Translation (MT)

Translating text from one language to another is a challenging task, even for humans to try to fully capture the style and nuanced meaning of the original.

- While research focuses on trying to produce the fully automatic, high-quality translation, there are many tasks for which a rough translation is sufficient.
- The differences between languages include systematic differences that can be modeled in some way and idiosyncratic and lexical differences that must be dealt with one by one.

Machine translation focuses on:

- **Faithfulness:** The meaning of the text has been preserved.
- **Fluency:** The translated text sounds natural to a native speaker and also maintains the style of the original text.

Why MT Is Hard

Given the Japanese phrase:

- *fukaku hansei shite orimasu*

If this is translated to English as:

- *we apologize,* it is not faithful to the original meaning

But, if we translate it as:

- *we are deeply reflecting (on our past behavior, and what we did wrong, and how to avoid the problem next time),* the translation is not fluent.

Example from Jurafsky and Martin text

Differences Between Languages

Morphological differences

- Number of morphemes per word and sentence
  - Isolating languages: Vietnamese and Cantonese, each word has one morpheme
  - Polysynthetic languages: “Eskimo,” a single word has many morphemes corresponding to a complete sentence
- Degree to which morphemes are segmentable
  - Agglutinative, morphemes have clean boundaries (Turkish)
  - Fusion languages, single affix may have multiple morphemes (Russian)

Differences Between Languages

Syntactic differences

- Basic word order of verbs, subjects, and objects
  - SVO: English, Mandarin, French, German, …
  - SOV: Hindi, Japanese
  - VSO: Classical Arabic and Biblical Hebrew
- Head-marking and dependent-marking languages
  - Mark relation between dependent and head on the head
    - English marks possessive on dependent: *the man's house*
    - Hungarian marks possessive on the head noun (Hungarian equivalent of): *the man house-his*

Differences Between Languages

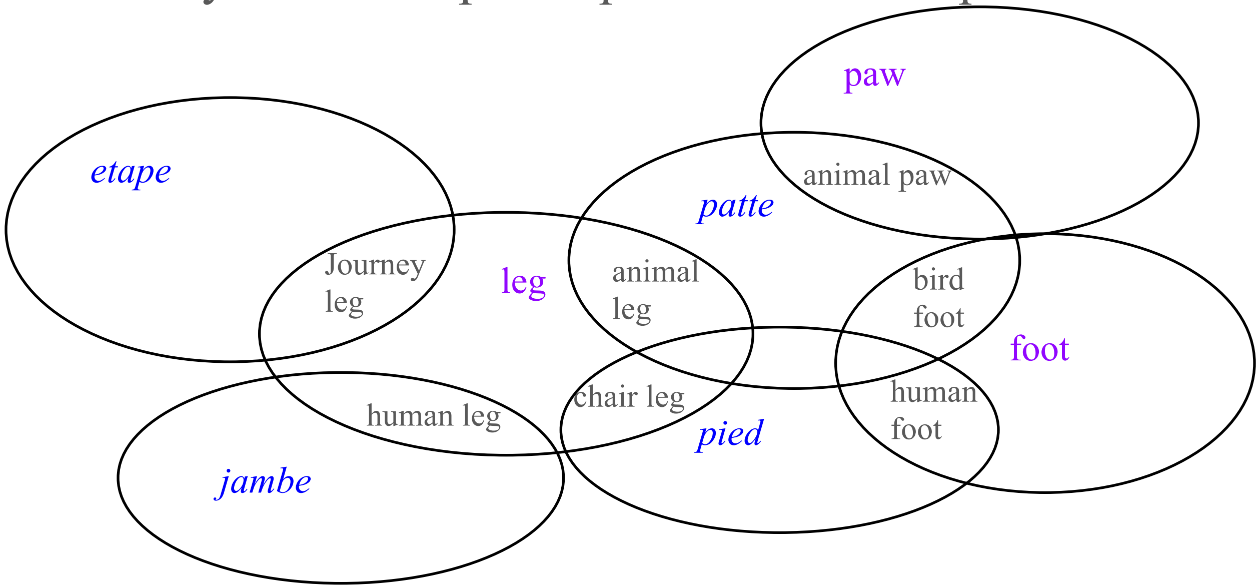
Syntactic differences

- Direction of motion with respect to verb
  - English direction on particle: *the bottle floated out*
  - Spanish direction on verb: *la botella salio' flotando*
- Grammatical constraints on matching gender-marked words
- Many others…

Differences Between Languages

Semantic differences

- Lexical gap
  - One language doesn't have a word for concept in another
- Differences in way that conceptual space is divided up for different words



The complex overlap between English leg, foot, etc. and various French translations. *(Jurafsky & Martin, Figure 21.2)*

Machine Translation Task

School of Information Studies
Syracuse University

Classical MT/Machine Translation

In this line of MT research, approaches can be classified according to the level of unit of translation.

- Utilizes word translation dictionaries
- Direct translation uses a word translation approach
- Transfer approaches use syntactic phrase and semantic units as the unit of translation

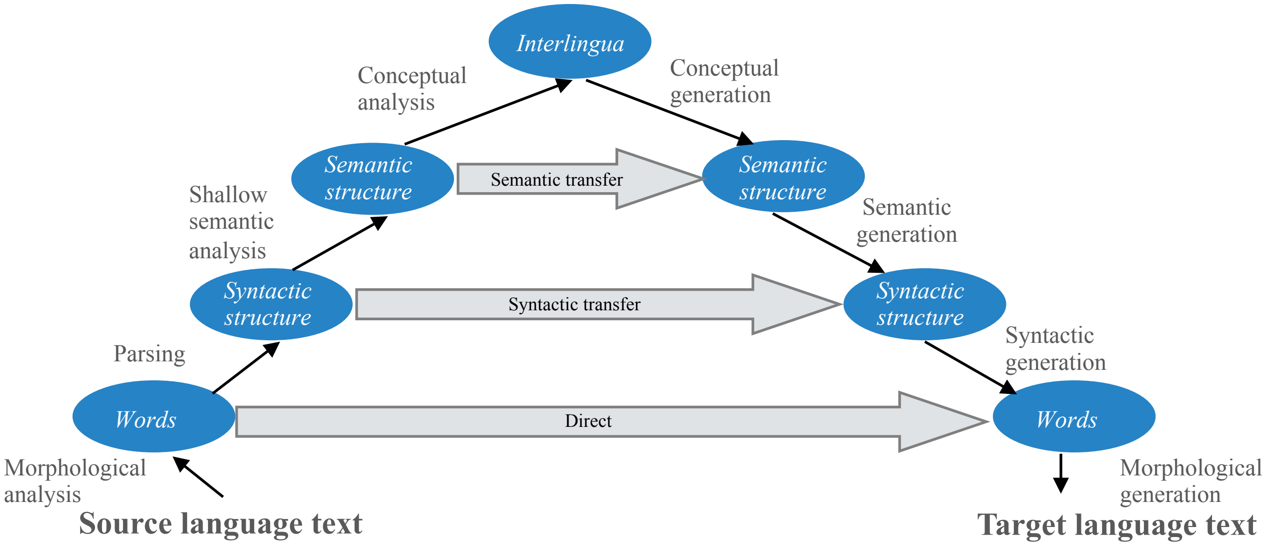


Figure 25.3 The Vauquois triangle

Statistical Approaches

Build probabilistic models of faithfulness and fluency and combine the models to get the most probable translation

Modeled as a noisy channel

- *“Pretend that the foreign input F is a corrupted version of the target language output E and the task is to discover the hidden sentence E that generated the observed sentence F.”*
- Informally, we refer to translating from French to English.

Statistical Approaches

Requires two models

- **Language model** to compute *P*(E), probability that any sequence E of English words is a sentence
  - Language model gives fluency
- **Translation model** to compute *P*(F|E), conditional probability that French sentence F was a translation of English sentence E
  - Translation model gives faithfulness
- **Given French sentence f, its English translation e** is

  arg max (all e in E) *P*(e) * *P*(f|e)

  - Note that this appears backwards to translate from French to English, but we invoke Bayes' theorem to define the translator.

Language Models for Fluency

Language model to compute *P*(E), the probability that a sequence of words is in the language

This models fluency, how likely it is that an English language speaker would say this sentence

In practice, learn probabilities of bigrams in the language to be translated from instead of entire sentences

- Translation has improved greatly due to large corpora
  - See Google Translate

Translation Models for Faithfulness

Translation model to compute *P*(F|E), the probability that the French sentence f came from the English sentence e

- Learn probabilities from parallel corpora
- Model the translation as word translation combined with alignment probabilities
  - E: *And the program has been implemented.*
  - F: *Le programme a ete mis en application.*
  - Alignment variables: (2, 3, 4, 5, 6, 6, 6) gives

Le	->	*the*
Programme	->	*program*
a	->	*has*
ete	->	*been*
mis	->	*implemented*
en	->	*implemented*
application	->	*implemented*

Alignment and Parallel Corpora

The translation model uses probabilities of word alignment.

Word alignment models are automatically trained from parallel corpora where each document is given in two or more languages.

- Hansard Corpus
  - Canadian parliament documents for French, English, and a variety of native American languages
- United Nations proceedings documents
- LDC (Linguistic Data Consortium) has corpora in several language pairs

Literary parallel corpora are not as suitable because of the stronger presence of literary devices, such as metaphor.

MT Evaluation

Human raters can evaluate along the two dimensions of fluency and fidelity (and there are several individual metrics for each of these dimensions)

BLEU automatic evaluation system

- Evaluation corpus contains human-generated translations
- Metrics evaluate how closely the system-generated translations correspond to the human ones
  - Measure the number of words and phrases overlapping between the human translation and the machine translation

Introduction to Summarization

School of Information Studies
Syracuse University

Summarization

Text summarization is the process of distilling the most important information from a text to produce an abridged version for a particular task and user.

- Definition adapted from Mani and Maybury, 1999

Types of summaries in current research

- Outlines or abstracts of any document, article, etc.
- Snippets summarizing a webpage or a search engine results page
- Action items or other summaries of a business meeting
- Summaries of email threads
- Simplifying text by compressing sentences
- Keyword extraction

Single vs. Multiple Documents

Single-document summarization

- Given a single document, produce a gist of the content in the form of an abstract or outline

Multiple-document summarization

- Given a group of documents, produce a gist of the content and create a cohesive answer that combines information from each document
  - A series of news stories on the same event
  - A set of webpages about some topic or question

Extractive vs. Abstractive

Abstractive summarization

- Express the ideas in the source documents using (at least in part) different words
- How humans typically approach summarization

Extractive summarization

- Create the summary from phrases or sentences in the source document(s)

Example of Extractive Summary

This example summary is produced by an online demo site.

On the next slide, about half the news article is shown.

- Summary does not paraphrase any sentences
- Extracted sentences occur primarily near the beginning of the article

More generally, paraphrasing may be needed to correct dangling references.

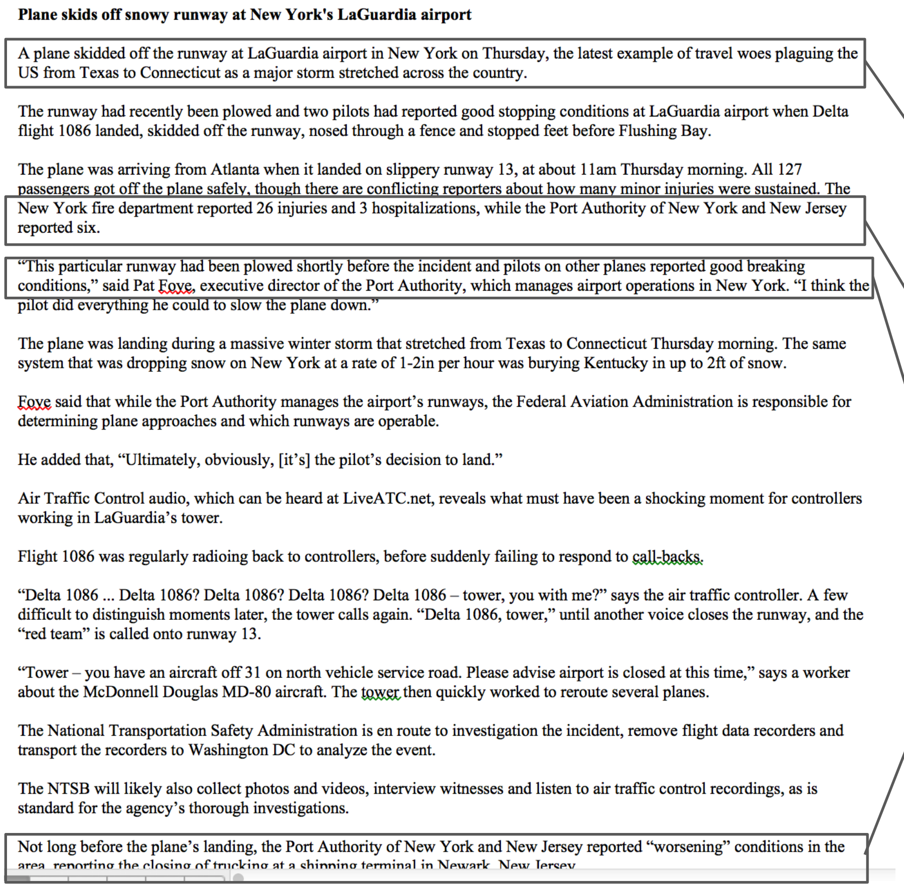
- Done from coreference resolution

Plane skids off snowy runway at New York's LaGuardia airport

A plane skidded off the runway at LaGuardia airport in New York on Thursday, the latest example of travel woes plaguing the US from Texas to Connecticut as a major storm stretched across the country. ***The New York fire department reported 26 injuries and 3 hospitalizations, while the Port Authority of New York and New Jersey reported six.***

"This particular runway had been plowed shortly before the incident and pilots on other planes reported good breaking conditions," said Pat Foye, executive director of the Port Authority, which manages airport operations in New York. Not long before the plane's landing, the Port Authority of New York and New Jersey reported "worsening" conditions in the area, reporting the closing of trucking at a shipping terminal in Newark, New Jersey.

Example of Extractive Summary



Plane skids off snowy runway at New York's LaGuardia airport

A plane skidded off the runway at LaGuardia airport in New York on Thursday, the latest example of travel woes plaguing the US from Texas to Connecticut as a major storm stretched across the country.

The runway had recently been plowed and two pilots had reported good stopping conditions at LaGuardia airport when Delta flight 1086 landed, skidded off the runway, nosed through a fence and stopped feet before Flushing Bay.

The plane was arriving from Atlanta when it landed on slippery runway 13, at about 11am Thursday morning. All 127 passengers got off the plane safely, though there are conflicting reporters about how many minor injuries were sustained. The New York fire department reported 26 injuries and 3 hospitalizations, while the Port Authority of New York and New Jersey reported six.

"This particular runway had been plowed shortly before the incident and pilots on other planes reported good breaking conditions," said Pat Foye, executive director of the Port Authority, which manages airport operations in New York. "I think the pilot did everything he could to slow the plane down."

The plane was landing during a massive winter storm that stretched from Texas to Connecticut Thursday morning. The same system that was dropping snow on New York at a rate of 1-2in per hour was burying Kentucky in up to 2ft of snow.

Foye said that while the Port Authority manages the airport's runways, the Federal Aviation Administration is responsible for determining plane approaches and which runways are operable.

He added that, "Ultimately, obviously, [it's] the pilot's decision to land."

Air Traffic Control audio, which can be heard at LiveATC.net, reveals what must have been a shocking moment for controllers working in LaGuardia's tower.

Flight 1086 was regularly radioing back to controllers, before suddenly failing to respond to call-backs.

"Delta 1086 ... Delta 1086? Delta 1086? Delta 1086? Delta 1086 – tower, you with me?" says the air traffic controller. A few difficult to distinguish moments later, the tower calls again. "Delta 1086, tower," until another voice closes the runway, and the "red team" is called onto runway 13.

"Tower – you have an aircraft off 31 on north vehicle service road. Please advise airport is closed at this time," says a worker about the McDonnell Douglas MD-80 aircraft. The tower then quickly worked to reroute several planes.

The National Transportation Safety Administration is en route to investigation the incident, remove flight data recorders and transport the recorders to Washington DC to analyze the event.

The NTSB will likely also collect photos and videos, interview witnesses and listen to air traffic control recordings, as is standard for the agency's thorough investigations.

Not long before the plane's landing, the Port Authority of New York and New Jersey reported "worsening" conditions in the area, reporting the closing of trucking at a shipping terminal in Newark, New Jersey.

Plane skids off snowy runway at New York's LaGuardia airport

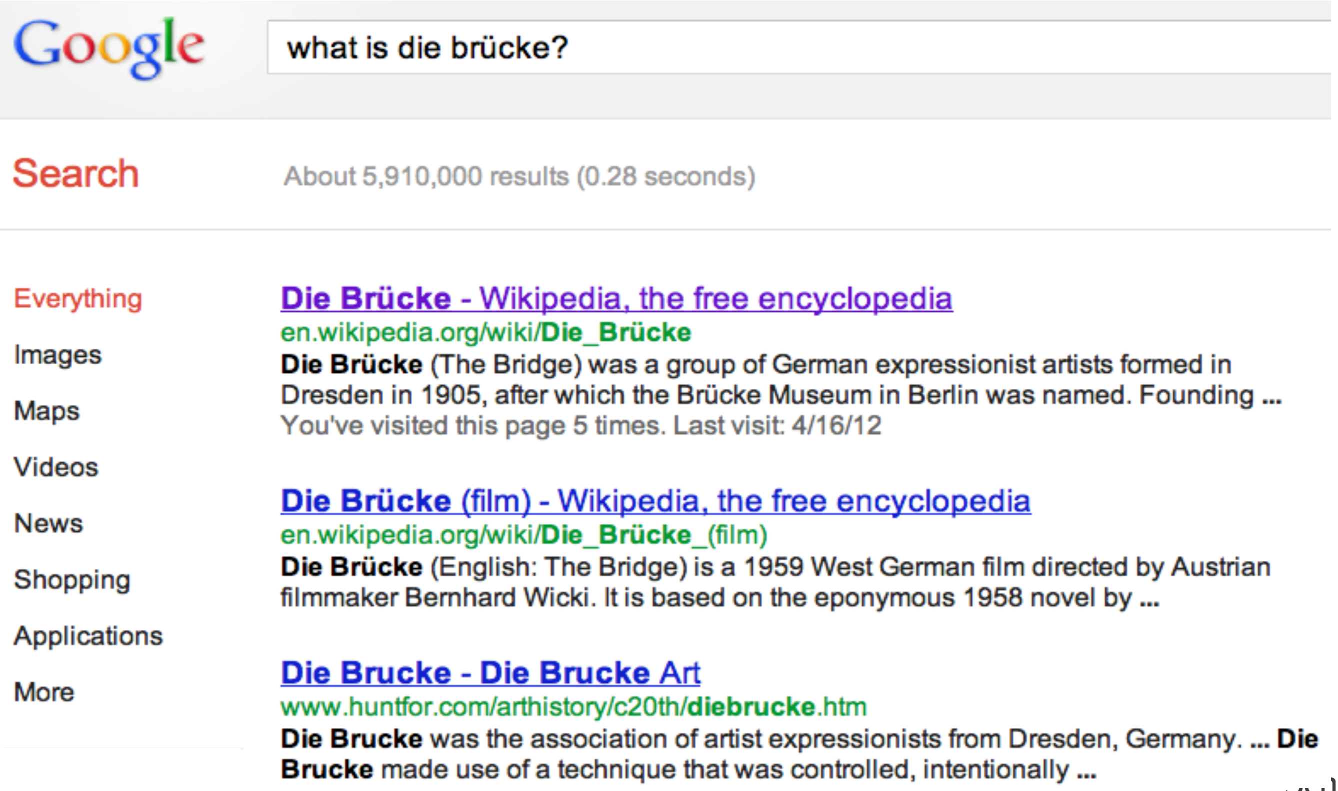
A plane skidded off the runway at LaGuardia airport in New York on Thursday, the latest example of travel woes plaguing the US from Texas to Connecticut as a major storm stretched across the country. ***The New York fire department reported 26 injuries and 3 hospitalizations, while the Port Authority of New York and New Jersey reported six.***

"This particular runway had been plowed shortly before the incident and pilots on other planes reported good breaking conditions," said Pat Foye, executive director of the Port Authority, which manages airport operations in New York. Not long before the plane's landing, the Port Authority of New York and New Jersey reported "worsening" conditions in the area, reporting the closing of trucking at a shipping terminal in Newark, New Jersey.

Summarization for Web Search: Snippets

Create **snippets** summarizing a webpage for a query.

- Google: 156 characters (about 26 words) plus title and link



what is die brücke?

About 5,910,000 results (0.28 seconds)

Die Brücke - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Die_Brücke**
Die Brücke (The Bridge) was a group of German expressionist artists formed in Dresden in 1905, after which the Brücke Museum in Berlin was named. Founding ...
You've visited this page 5 times. Last visit: 4/16/12

Die Brücke (film) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Die_Brücke**_(film)
Die Brücke (English: The Bridge) is a 1959 West German film directed by Austrian filmmaker Bernhard Wicki. It is based on the eponymous 1958 novel by ...

Die Brucke - **Die Brucke** Art
www.huntfor.com/arthistory/c20th/**diebrucke**.htm
Die Brucke was the association of artist expressionists from Dresden, Germany. ... **Die Brucke** made use of a technique that was controlled, intentionally ...

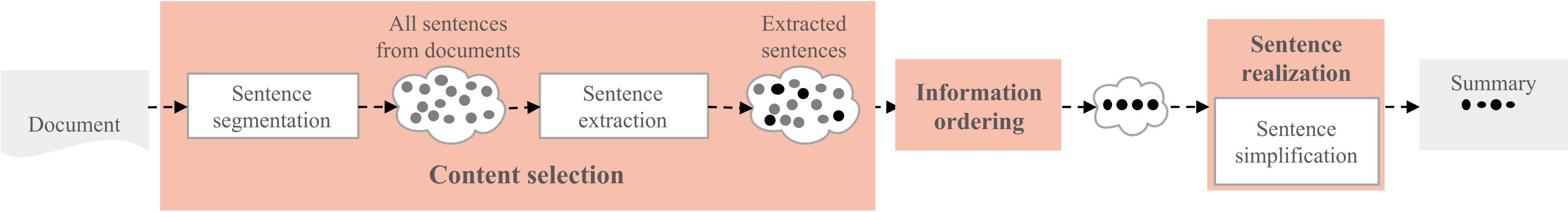
Summarization Task

School of Information Studies
Syracuse University

Summarization Typical Approaches

Currently, achieve extraction instead of a true rephrasing.

- **Content selection**
  - Identify the sentences or clauses to extract
- **Information ordering**
  - How to order the selected units
- **Sentence realization**
  - Perform cleanup on the extracted units so that they are fluent in their new context
    - Example: replacing pronoun or other references left dangling



Content Selection: Centrality Methods

Centrality methods select sentences based on properties of words and sentences.

- Don’t require annotated training data
- But do draw on semantics of a background corpus

The **simple approach** is to select sentences that have more informative words according to saliency defined from a topic signature of the document.

- Topics are learned from the larger corpus
- A topic is a set of words grouped together that score high on relatedness, using mutual information or other semantic similarity

Centroid-based summarization uses log-likelihood ratios for words, computing the probability of observing the word in the input more often than in the background corpus.

Content Selection: TextRank

The **TextRank or LexRank method** scores the importance of sentences in a similar way to PageRank on webpages.

- Each sentence is a vertex of a graph and the PageRank algorithm assigns scores based on the similarity of the words in the sentences.
- The PageRank algorithm comes from search engines where pages are scored based on link structure.
- TextRank also uses semantic similarity measures, which may be learned from the background corpus.

Content Selection: Other Methods

Methods based on rhetorical parsing use coherence relations to identify satellite and nucleus sentences.

Machine learning methods score sentences for importance from a corpus of sentences assigned importance scores.

- Use features based on:
  - Position
  - Cue phrases
  - Word informativeness
  - Sentence length
  - Cohesion (computing lexical chains of the document)

Information Ordering

Single documents can simply keep the document ordering.

Chronological ordering for multiple documents

- Order sentences by the date of the document (for summarizing news).

Coherence ordering

- Choose orderings that make neighboring sentences similar.
- Choose orderings in which neighboring sentences discuss the same entity.

Topical ordering

- Learn the ordering of topics in the source documents.

Simplifying Sentences

Simplest method: Parse sentences and use rules to decide which modifiers to prune.

- More recently, a wide variety of machine learning methods

Appositives	Rajam, ~~28, an artist who was living at the time in Philadelphia~~, found the inspiration in the back of city magazines.
Attribution clauses	Rebels agreed to talks with government officials, ~~international observers said Tuesday.~~
Prepositional phrases without named entities	The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases [~~PP to a sustainable number~~]]
Initial adverbials	“~~For example~~”, “~~On the other hand~~”, “~~As a matter of fact~~”, “~~At this point~~”

Summarization Evaluation

Extrinsic (task-based) evaluation: humans are asked to rate the summaries according to how well they are enabled to perform a specific task

Intrinsic (task-independent) evaluation

- Human judgments to rate the summaries
- ROUGE (recall oriented understudy gisting evaluation)
  - Humans generate summaries for a document collection.
  - System-generated summaries are rated according to how close they come to the human-generated summary.
  - Measures have included unigram overlap, bigram overlap, and longest common subsequence.
- Pyramid method
  - Humans identify "units of meaning," which are short phrases, and then an overlap measure is computed.