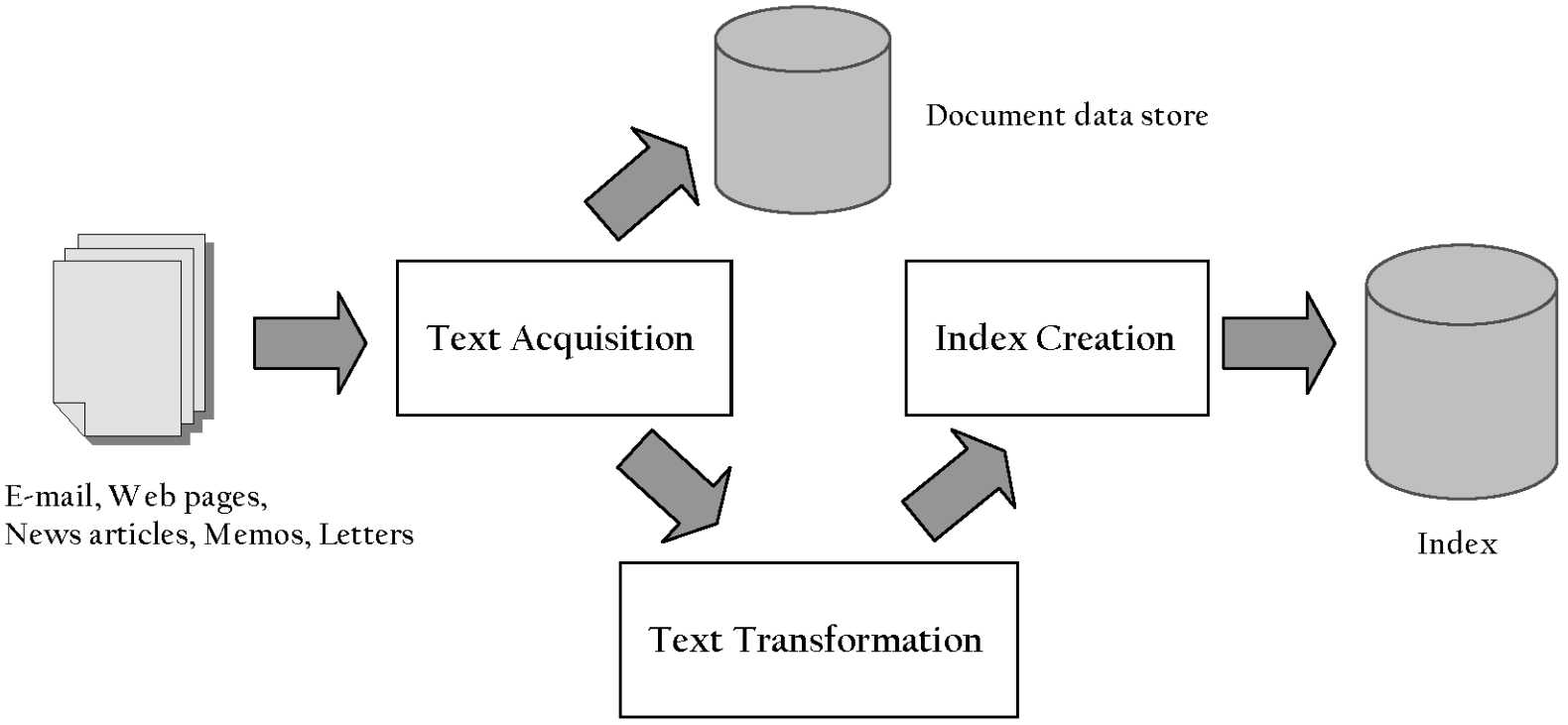
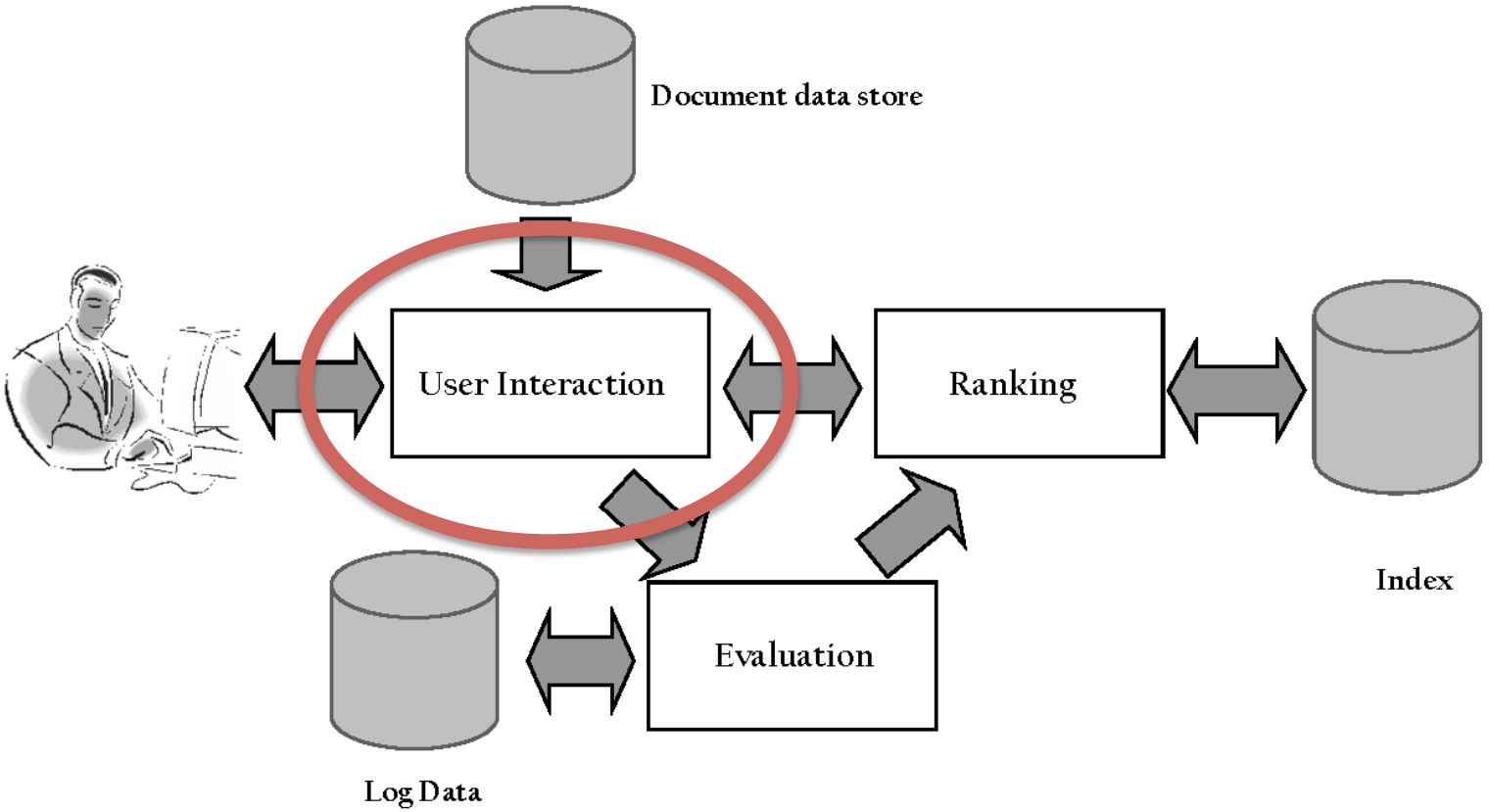
IS4200/CS6200 Information Retrieval

David Smith
College of Computer and Information Science
Northeastern University

Previously: Indexing Process



Query Process



Information Needs

- An *information need* is the underlying cause of the query that a person submits to a search engine
  - sometimes called *query intent*
- Categorized using variety of dimensions
  - e.g., number of relevant documents being sought
  - type of information that is needed
  - type of task that led to the requirement for information

Queries and Information Needs

- A query can represent very different information needs
  - May require different search techniques and ranking algorithms to produce the best rankings
- A query can be a poor representation of the information need
  - User may find it difficult to express the information need
  - User is encouraged to enter short queries both by the search engine interface, and by the fact that long queries don't work

Interaction

- Interaction with the system occurs
  - during query formulation and reformulation
  - while browsing the result
- Key aspect of effective retrieval
  - users can't change ranking algorithm but can change results through interaction
  - helps refine description of information need
    - e.g., same initial query, different information needs
    - how does user describe what they don't know?

ASK Hypothesis

- Belkin et al (1982) proposed a model called Anomalous State of Knowledge
- ASK hypothesis:
  - difficult for people to define exactly what their information need is, because that information is a gap in their knowledge
  - Search engine should look for information that fills those gaps
- Interesting ideas, little practical impact (yet)

Keyword Queries

- Query languages in the past were designed for professional searchers (*intermediaries*)

User query:
Are there any cases which discuss negligent maintenance or failure to maintain aids to navigation such as lights, buoys, or channel markers?

Intermediary query:
NEGLECT! FAIL! NEGLIG! /5 MAINT! REPAIR! /P NAVIGAT! /5 AID EQUIP! LIGHT BUOY "CHANNEL MARKER"

Keyword Queries

- Simple, *natural language* queries were designed to enable everyone to search
- Current search engines do not perform well (in general) with natural language queries
- People trained (in effect) to use keywords
  - compare average of about 2.3 words/web query to average of 30 words/CQA query
- Keyword selection is not always easy
  - query refinement techniques can help

Query Reformulation

- Rewrite or transform original query to better match underlying intent
- Can happen implicitly or explicitly (suggestion)
- Many techniques
  - Query-based stemming
  - Spelling correction
  - Segmentation
  - Substitution
  - Expansion

Query-Based Stemming

- Make decision about stemming at query time rather than during indexing
  - improved flexibility, effectiveness
- Query is expanded using word variants
  - documents are not stemmed
  - e.g., “rock climbing” expanded with “climb”, not stemmed to “climb”

Stem Classes

- A *stem class* is the group of words that will be transformed into the same stem by the stemming algorithm
  - generated by running stemmer on large corpus
  - e.g., Porter stemmer on TREC News

/bank banked banking bankings banks
/ocean oceaneering oceanic oceanics oceanization oceans
/polic polical polically police policeable policed
-policement policer policers polices policial
-policically policier policiers policies policing
-policization policize policly policy policying policys

Stem Classes

- Stem classes are often too big and inaccurate
- Modify using analysis of *word co-occurrence*
- *Assumption:*
  - Word variants that could substitute for each other should co-occur often in documents
    - e.g., reduces previous example /polic and /bank classes to

      /policies policy
      /police policed policing
      /bank banking banks

Query Log

- Records all queries and documents clicked on by users, along with timestamp
- Used heavily for query transformation, query suggestion
- Also used for query-based stemming
  - Word variants that co-occur with other query words can be added to query
    - e.g., for the query “tropical fish”, “fishes” may be found with “tropical” in query log, but not “fishing”
    - Classic example: “strong tea” *not* “powerful tea”

Modifying Stem Classes

1. For all pairs of words in the stem classes, count how often they co-occur in text windows of W words. W is typically in the range 50-100.

2. Compute a co-occurrence or association metric for each pair. This measures how strong the association is between the words.

3. Construct a graph where the vertices represent words and the edges are between words whose co-occurrence metric is above a threshold T.

4. Find the connected components of this graph. These are the new stem classes.

Modifying Stem Classes

- Dices' Coefficient is an example of a term association measure
  - $2.n_{ab}/(n_a + n_b)$
  - where n_x is the number of windows containing *x*
- Two vertices are in the same connected component of a graph if there is a path between them
  - forms word *clusters*
- Example output of modification
- When would this fail?

/policies policy
/police policed policing
/bank banking banks

Query Segmentation

- Break up queries into important "chunks"
  - e.g., "new york times square" becomes "new york" "times square"
- Possible approaches:

Treat each term as a concept

[members] [rock] [group] [nirvana]

Treat every adjacent pair of terms as a concept

[members rock] [rock group] [group nirvana]

Treat all terms within a noun phrase "chunk" as a concept

[members] [rock group nirvana]

Treat all terms that occur in common queries as a single concept

[members] [rock group] [nirvana]

The Thesaurus

- Used in early search engines as a tool for *indexing* and *query formulation*
  - specified preferred terms and relationships between them
  - also called *controlled vocabulary*
  - or *authority list*
- Particularly useful for *query expansion*
  - adding synonyms or more specific terms using query operators based on thesaurus
  - improves search effectiveness

MeSH Thesaurus

MeSH Heading	**Neck Pain**
Tree Number	C10.597.617.576
Tree Number	C23.888.592.612.553
Tree Number	C23.888.646.501
Entry Term	Cervical Pain
Entry Term	Neckache
Entry Term	Anterior Cervical Pain
Entry Term	Anterior Neck Pain
Entry Term	Cervicalgia
Entry Term	Cervicodynia
Entry Term	Neck Ache
Entry Term	Posterior Cervical Pain
Entry Term	Posterior Neck Pain

Query Expansion

- A variety of *automatic* or *semi-automatic* query expansion techniques have been developed
  - goal is to improve effectiveness by matching related terms
  - semi-automatic techniques require user interaction to select best expansion terms
- Query suggestion is a related technique
  - alternative queries, not necessarily more terms

Query Expansion

- Approaches usually based on an analysis of term co-occurrence
  - either in the entire document collection, a large collection of queries, or the top-ranked documents in a result list
  - query-based stemming also an expansion technique
- Automatic expansion based on general thesaurus not effective
  - does not take context into account

Term Association Measures

- *Dice's Coefficient*

$$\frac{2.n_{ab}}{n_a+n_b} \overset{rank}{=} \frac{n_{ab}}{n_a+n_b}$$

- *Mutual Information*

$$\log \frac{P(a,b)}{P(a)P(b)} = \log N.\frac{n_{ab}}{n_a.n_b} \overset{rank}{=} \frac{n_{ab}}{n_a.n_b}$$

Term Association Measures

- Mutual Information measure favors low frequency terms
- *Expected Mutual Information Measure* (EMIM)

$$P(a,b).\log\frac{P(a,b)}{P(a)P(b)} = \frac{n_{ab}}{N}\log(N.\frac{n_{ab}}{n_a.n_b}) \overset{rank}{=} n_{ab}.\log(N.\frac{n_{ab}}{n_a.n_b})$$

  - actually only 1 part of full EMIM, focused on word occurrence

Term Association Measures

- *Pearson's Chi-squared (χ^2) measure*
  - compares the number of co-occurrences of two words with the expected number of co-occurrences if the two words were independent
  - normalizes this comparison by the expected number
  - also limited form focused on word co-occurrence

$$\frac{(n_{ab} - N.\frac{n_a}{N}.\frac{n_b}{N})^2}{N.\frac{n_a}{N}.\frac{n_b}{N}} \underline{=}^{rank} \frac{(n_{ab} - \frac{1}{N}.n_a.n_b)^2}{n_a.n_b}$$

Association Measure Summary

Measure	*Formula*
Mutual information (MIM)	$\frac{n_{ab}}{n_a.n_b}$
Expected Mutual Information ($EMIM$)	$n_{ab}.\log(N.\frac{n_{ab}}{n_a.n_b})$
Chi-square (χ^2)	$\frac{(n_{ab}-\frac{1}{N}.n_a.n_b)^2}{n_a.n_b}$
Dice's coefficient ($Dice$)	$\frac{n_{ab}}{n_a+n_b}$

Association Measure Example

MIM	*EMIM*	χ^2	*Dice*
trmm	forest	trmm	forest
itto	tree	itto	exotic
ortuno	rain	ortuno	timber
kuroshio	island	kuroshio	rain
ivirgarzama	like	ivirgarzama	banana
biofunction	fish	biofunction	deforestation
kapiolani	most	kapiolani	plantation
bstilla	water	bstilla	coconut
almagreb	fruit	almagreb	jungle
jackfruit	area	jackfruit	tree
adeo	world	adeo	rainforest
xishuangbanna	america	xishuangbanna	palm
frangipani	some	frangipani	hardwood
yuca	live	yuca	greenhouse
anthurium	plant	anthurium	logging

Most strongly associated words for “tropical” in a collection of TREC news stories. Co-occurrence counts are measured at the document level.

Association Measure Example

MIM	*EMIM*	χ^2	*Dice*
zoologico	water	arlsq	species
zapanta	species	happyman	wildlife
wrint	wildlife	outerlimit	fishery
wpfmc	fishery	sportk	water
weighout	sea	lingcod	fisherman
waterdog	fisherman	longfin	boat
longfin	boat	bontadelli	sea
veracruzana	area	sportfisher	habitat
ungutt	habitat	billfish	vessel
ulocentra	vessel	needlefish	marine
needlefish	marine	damaliscu	endanger
tunaboat	land	bontebok	conservation
tsolwana	river	taucher	river
olivacea	food	orangemouth	catch
motoroller	endanger	sheepshead	island

Most strongly associated words for “fish” in a collection of TREC news stories.

Association Measure Example

MIM	*EMIM*	χ^2	*Dice*
zapanta	wildlife	gefilte	wildlife
plar	vessel	mbmo	vessel
mbmo	boat	zapanta	boat
gefilte	fishery	plar	fishery
hapc	species	hapc	species
odfw	tuna	odfw	catch
southpoint	trout	southpoint	water
anadromous	fisherman	anadromous	sea
taiffe	salmon	taiffe	meat
mollie	catch	mollie	interior
frampton	nmf	frampton	fisherman
idfg	trawl	idfg	game
billingsgate	halibut	billingsgate	salmon
sealord	meat	sealord	tuna
longline	shellfish	longline	caught

Most strongly associated words for “fish” in a collection of TREC news stories. Co-occurrence counts are measured in windows of 5 words.

Association Measures

- Associated words are of little use for expanding the query “tropical fish”
- Expansion based on whole query takes context into account
  - e.g., using Dice with term “tropical fish” gives the following highly associated words:

    goldfish, reptile, aquarium, coral, frog, exotic, stripe, regent, pet, wet
- Impractical for all possible queries, other approaches used to achieve this effect

Other Approaches

- Pseudo-relevance feedback
  - expansion terms based on top retrieved documents for initial query
- Context vectors
  - Represent words by the words that co-occur with them
    - e.g., top 35 most strongly associated words for “aquarium” (using Dice’s coefficient):

      zoology, cranmore, jouett, zoo, goldfish, fish, cannery, urchin, reptile, coral, animal, mollusk, marine, underwater, plankton, mussel, oceanography, mammal, species, exhibit, swim, biologist, cabrillo, saltwater, creature, reef, whale, oceanic, scuba, kelp, invertebrate, park, crustacean, wild, tropical
  - Rank words for a query by ranking context vectors

Other Approaches

- Query logs
  - Best source of information about queries and related terms
    - short pieces of text and click data
  - e.g., most frequent words in queries containing "tropical fish" from MSN log:

    stores, pictures, live, sale, types, clipart, blue, freshwater, aquarium, supplies
  - Query suggestion based on finding similar queries
    - group based on click data
  - Query reformulation/expansion based on term associations in logs

Query Suggestion using Logs

Orig. Query	NDCG@10	New Query	NDCG@10
moths	0.1714	where do you find white moths	0.389
iron	0.0	normal iron levels for women	0.7315
getting organized	0.2341	free printable planner	0.4603
arizona game and fish	0.164	az fish and game	0.1712
kcs	0.0	kansas city southern	0.4026
starbucks	0.5529	sbux	0.6828
used car parts	0.1197	used auto parts	0.3083
used car parts	0.1197	salvage yards	0.3464
dinosaurs	0.0	dinosaurs pictures	0.224
map	0.0	www.mapquest.com	0.0792

Orig. Query	NDCG@10	New Query	NDCG@10
penguins	0.2578	official pittsburgh penguins website	0.5062
bellevue	0.103	bellevue washington	0.6823
tornadoes	0.468	questions and answers about tornadoes	0.7382
ocd	0.041	obsessive compulsive disorder	0.2107
kcs	0.0	kansas city southern	0.4026
kcs	0.0	www kcsi com	0.5049
air travel information	0.0	permitted and prohibited items	0.0245
atari	0.1821	infogrames	0.336
iron	0.0	fe	0.359
tornadoes	0.468	noaa tornadoes	0.6153

Query Reformulation using Logs

	Original Query	Expanded Query	Original Query	Expanded Query
MSN Log	hunting deaths	hunting #syn(deaths accidents)	railway accidents	#syn(railway train) accidents
	new fuel sources	new #syn(fuel energy) sources	oscar winner selection	oscar winner #syn(selection promotion)
	educational standards	#syn(educational teaching) standards	marine vegetation	marine #syn(vegetation plants)
	automobile recalls	#syn(automobile auto) recalls	overseas tobacco sales	overseas #syn(tobacco cigarettes) sales
	doctor assisted suicides	#syn(doctor physicians) assisted suicides	food drug laws	food drug #syn(laws act)
	cheese production	cheese #syn(production companies)	volkswagen mexico	#syn(volkswagen vw) mexico
	illegal immigrant wages	illegal immigrant #syn(wages working)	chevrolet trucks	#syn(chevrolet chevy) trucks
Anchor Log	hunting deaths	hunting #syn(deaths accidents)	railway accidents	#syn(railway railroad) accidents
	new fuel sources	new #syn(fuel energy) sources	pearl farming	pearl #syn(farming industry)
	educational standards	#syn(educational teaching) standards	eskimo history eskimo	#syn(history culture)
	automobile recalls	#syn(automobile auto)	international art crime	international art #syn(crime fraud)
	doctor assisted suicides	#syn(doctor physicians) assisted suicides	wildlife extinction	#syn(wildlife animals) extinction
	cheese production	cheese #syn(production prices)	blood alcohol fatalities	blood alcohol #syn(fatalities deaths)
	illegal immigrant wages	illegal #syn(immigrant workers) wages	windmill electricity	windmill #syn(electricity power)

Spell Checking

- Important part of query processing
  - 10-15% of all web queries have spelling errors
- Errors include typical word processing errors but also many other types, e.g.

poiner sisters
brimingham news
catamarn sailing
hair extenssions
marshmellow world
miniture golf courses
psyhics
home doceration

realstateisting.bc.com
akia 1080i manunal
ultimatwarcade
mainscourcebank
dellottitouche

Spell Checking

- Basic approach: suggest corrections for words not found in *spelling dictionary*
- Suggestions found by comparing word to words in dictionary using similarity measure
- Most common similarity measure is *edit distance*
  - number of operations required to transform one word into the other

Edit Distance

- *Damerau-Levenshtein* distance
  - counts the minimum number of insertions, deletions, substitutions, or transpositions of single characters required
  - e.g., Damerau-Levenshtein distance 1

    extenssions $\rightarrow$ extensions(insertion error)
    poiner $\rightarrow$ pointer (deletion error)
    marshmellow $\rightarrow$ marshmallow $(\text{substitution error})$
    brimingham $\rightarrow$ birmingham $(\text{transposition error})$
  - distance 2

    doceration $\rightarrow$ deceration
    deceration $\rightarrow$ decoration

Edit Distance

- Number of techniques used to speed up calculation of edit distances
  - restrict to words starting with same character
  - restrict to words of same or similar length
  - restrict to words that sound the same
- Last option uses a *phonetic code* to group words
  - e.g. Soundex

Soundex Code

1. Keep the first letter (in upper case).
2. Replace these letters with hyphens: a,e,i,o,u,y,h,w.
3. Replace the other letters by numbers as follows:

   1: b,f,p,v
   2: c,g,j,k,q,s,x,z
   3: d,t
   4: l
   5: m,n
   6: r

4. Delete adjacent repeats of a number.
5. Delete the hyphens.
6. Keep the first three numbers or pad out with zeros.

extenssions $\longrightarrow$ E235; extensions $\longrightarrow$ E235
marshmellow $\longrightarrow$ M625; marshmallow $\longrightarrow$ M625
brimingham $\longrightarrow$ B655; birmingham $\longrightarrow$ B655
poiner $\longrightarrow$ P560; pointer $\longrightarrow$ P536

Spelling Correction Issues

- Ranking corrections
  - "Did you mean..." feature requires accurate ranking of possible corrections
- Context
  - Choosing right suggestion depends on context (other words)
  - e.g., *lawers → lowers, lawyers, layers, lasers, lagers* but *trial lawers → trial lawyers*
- Run-on errors
  - e.g., "mainscourcebank"
  - missing spaces can be considered another single character error in right framework

Noisy Channel Model

- User chooses word *w* based on probability distribution *P(w)*
  - called the *language model*
  - can capture context information, e.g. $P(w_1|w_2)$
- User writes word, but noisy channel causes word *e* to be written instead with probability *P(e|w)*
  - called *error model*
  - represents information about the frequency of spelling errors

Noisy Channel Model

- Need to estimate probability of correction
  - $P(w|e) = P(e|w)P(w)$
- Estimate language model using context
  - e.g., $P(w) = \lambda P(w) + (1 - \lambda)P(w|w_p)$
  - w_p is previous word
- e.g.,
  - "fish tink"
  - "tank" and "think" both likely corrections, but $P(\text{tank}|\text{fish}) > P(\text{think}|\text{fish})$

Noisy Channel Model

- Language model probabilities estimated using corpus and query log
- Both simple and complex methods have been used for estimating error model
  - simple approach: assume all words with same edit distance have same probability, only edit distance 1 and 2 considered
  - more complex approach: incorporate estimates based on common typing errors

Example Spellcheck Process

1. Tokenize the query.
2. For each token, a set of alternative words and pairs of words is found using an edit distance modified by weighting certain types of errors as described above. The data structure that is searched for the alternatives contains words and pairs from both the query log and the trusted dictionary.
3. The noisy channel model is then used to select the best correction.
4. The process of looking for alternatives and finding the best correction is repeated until no better correction is found.

e.g.,
miniture golfcurses
miniature golfcourses
miniature golf courses

Relevance Feedback

- User identifies relevant (and maybe non-relevant) documents in the initial result list
- System modifies query using terms from those documents and reranks documents
  - example of simple machine learning algorithm using training data
  - but, very little training data
- Pseudo-relevance feedback just assumes top-ranked documents are relevant – no user input
  - In machine learning, aka self-training or bootstrapping

Relevance Feedback Example

1. Badmans **Tropical Fish**

   A freshwater aquarium page covering all aspects of the **tropical fish** hobby. ... to Badman's **Tropical Fish**. ... world of aquariology with Badman's **Tropical Fish**. ...

2. **Tropical Fish**

   Notes on a few species and a gallery of photos of African cichlids.

3. The **Tropical** Tank Homepage - **Tropical Fish** and Aquariums

   Info on **tropical fish** and **tropical** aquariums, large **fish** species index with ... Here you will find lots of information on **Tropical Fish** and Aquariums. ...

4. **Tropical Fish** Centre

   Offers a range of aquarium products, advice on choosing species, feeding, and health care, and a discussion board.

5. **Tropical fish** - Wikipedia, the free encyclopedia

   Tropical fish are popular aquarium **fish** , due to their often bright coloration. ... Practical Fishkeeping • **Tropical Fish** Hobbyist • Koi. Aquarium related companies: ...

6. **Tropical Fish** Find

   Home page for **Tropical Fish** Internet Directory ... stores, forums, clubs, **fish** facts, **tropical fish** compatibility and aquarium ...

7. Breeding **tropical fish**

   ... intrested in keeping and/or breeding **Tropical**, Marine, Pond and Coldwater **fish**. ... Breeding **Tropical Fish** ... breeding **tropical**, marine, coldwater & pond **fish**. ...

8. FishLore

   Includes **tropical** freshwater aquarium how-to guides, FAQs, **fish** profiles, articles, and forums.

9. Cathy's **Tropical Fish** Keeping

   Information on setting up and maintaining a successful freshwater aquarium.

10. **Tropical Fish** Place

    Tropical Fish information for your freshwater **fish** tank ... great amount of information about a great hobby, a freshwater **tropical fish** tank. ...

Top 10 documents for “tropical fish”

Relevance Feedback Example

- If we assume top 10 are relevant, most frequent terms are (with frequency):

  a (926), td (535), href (495), http (357), width (345), com (343), nbsp (316), www (260), tr (239), htm (233), class (225), jpg (221)

  - too many stopwords and HTML expressions

- Use only snippets and remove stopwords

  tropical (26), fish (28), aquarium (8), freshwater (5), breeding (4), information (3), species (3), tank (2), Badman's (2), page (2), hobby (2), forums (2)

Relevance Feedback Example

- If document 7 ("Breeding tropical fish") is *explicitly* indicated to be relevant, the most frequent terms are:

  breeding (4), fish (4), tropical (4), marine (2), pond (2), coldwater (2), keeping (1), interested (1)

- Specific weights and scoring methods used for relevance feedback depend on retrieval model

Relevance Feedback

- Both relevance feedback and pseudo-relevance feedback are effective, but not used in many applications
  - pseudo-relevance feedback has reliability issues, especially with queries that don't retrieve many relevant documents
- Some applications use relevance feedback
  - filtering, "more like this"
- Query suggestion more popular
  - may be less accurate, but can work if initial query fails

Context and Personalization

- If a query has the same words as another query, should results be the same regardless of
  - who submitted the query,
  - why the query was submitted,
  - where the query was submitted, or
  - what other queries were submitted in the same session?
- These other factors (the *context*) could have a significant impact on relevance

User Models

- Generate user profiles based on documents that the person looks at
  - such as web pages visited, email messages, or word processing documents on the desktop
- Modify queries using words from profile
- Generally not effective
  - imprecise profiles, information needs can change significantly

Query Logs

- Query logs provide important contextual information that can be used effectively
- Context in this case is
  - previous queries that are the same
  - previous queries that are similar
  - query sessions including the same query
- Query history for individuals could be used for caching or query transformation

Local Search

- Location is context
- *Local search* uses geographic information to modify the ranking of search results
  - location derived from the query text
  - location of the device where the query originated
- e.g.,
  - “underworld 3 cape cod”
  - “underworld 3” from mobile device in Hyannis

Local Search

- Identify the geographic region associated with web pages
  - use location metadata that has been manually added to the document,
  - or identify locations such as place names, city names, or country names in text
- Identify the geographic region associated with the query
  - 10-15% of queries contain some location reference
- Rank web pages using location information in addition to text and link-based features

Extracting Location Information

- Type of information extraction
  - ambiguity and significance of locations are issues
- Location names are mapped to specific regions and coordinates

- Matching done by inclusion, distance

Snippet Generation

Tropical Fish

One of the U.K.s Leading suppliers of **Tropical**, Coldwater, Marine **Fish** and Invertebrates plus.. . next day **fish** delivery service ...

www.**tropicalfish**.org.uk/**tropical_fish**.htm Cached page

- Query-dependent document summary
- Simple summarization approach
  - rank each sentence in a document using a *significance factor*
  - select the top sentences for the summary
  - first proposed by Luhn in 50's

Sentence Selection

- Significance factor for a sentence is calculated based on the occurrence of *significant words*
  - If $f_{d,w}$ is the frequency of word *w* in document *d*, then *w* is a significant word if it is not a stopword and

$$f_{d,w} \geq \begin{cases} 7 - 0.1 \times (25 - s_d), & \text{if } s_d < 25 \\ 7, & \text{if } 25 \leq s_d \leq 40 \\ 7 + 0.1 \times (s_d - 40), & \text{otherwise} \end{cases}$$

    where s_d is the number of sentences in document *d*
  - text is *bracketed* by significant words (limit on number of non-significant words in bracket)

Sentence Selection

- Significance factor for bracketed text spans is computed by dividing the square of the number of significant words in the span by the total number of words
- e.g.,

w w w w w w w w w w w.
(Initial sentence)

w w s w s s w w s w w.
(Identify significant words)

w w [s w s s w w s] w w.
(Text span bracketed by significant words)

- Significance factor = $4^2/7 = 2.3$

Snippet Generation

- Involves more features than just significance factor
- e.g. for a news story, could use
  - whether the sentence is a heading
  - whether it is the first or second line of the document
  - the total number of query terms occurring in the sentence
  - the number of unique query terms in the sentence
  - the longest contiguous run of query words in the sentence
  - a density measure of query words (significance factor)
- Weighted combination of features used to rank sentences

Snippet Generation

- Web pages are less structured than news stories
  - can be difficult to find good summary sentences
- Snippet sentences are often selected from other sources
  - metadata associated with the web page
    - e.g., <meta name="description" content= ...>
  - external sources such as web directories
    - e.g., Open Directory Project, http://www.dmoz.org
- Snippets can be generated from text of pages like Wikipedia

Snippet Guidelines

- All query terms should appear in the summary, showing their relationship to the retrieved page
- When query terms are present in the title, they need not be repeated
  - allows snippets that do not contain query terms
- Highlight query terms in URLs
- Snippets should be readable text, not lists of keywords

Advertising

- *Sponsored search* – advertising presented with search results
- *Contextual advertising* – advertising presented when browsing web pages
- Both involve finding the most relevant advertisements in a database
  - An advertisement usually consists of a short text description and a link to a web page describing the product or service in more detail

Searching Advertisements

- Factors involved in ranking advertisements
  - similarity of text content to query
  - bids for keywords in query
  - popularity of advertisement
- Small amount of text in advertisement
  - dealing with vocabulary mismatch is important
  - expansion techniques are effective

Example Advertisements

fish tanks at Target
Find **fish tanks** Online. Shop & Save at Target.com Today.
www.target.com

Aquariums
540+ Aquariums at Great Prices.
fishbowls.pronto.com

Freshwater **Fish** Species
Everything you need to know to keep your setup clean and beautiful
www.FishChannel.com

Pet Supplies at Shop.com
Shop millions of products and buy from our trusted merchants.
shop.com

Custom **Fish Tanks**
Choose From 6,500+ Pet Supplies. Save On Custom **Fish Tanks**!
shopzilla.com

Advertisements retrieved for query “fish tank”

Searching Advertisements

- Pseudo-relevance feedback
  - expand query and/or document using the Web
  - use ad text or query for pseudo-relevance feedback
  - rank exact matches first, followed by stem matches, followed by expansion matches
- Query reformulation based on query log

Clustering Results

- Result lists often contain documents related to different *aspects* of the query topic
- *Clustering* is used to group related documents to simplify browsing

Example clusters for query “tropical fish”

Pictures (38)

Aquarium Fish (28)

Tropical Fish Aquarium (26)

Exporter (31)

Supplies (32)

Plants, Aquatic (18)

Fish Tank (15)

Breeding (16)

Marine Fish (16)

Aquaria (9)

Result List Example

1. Badmans **Tropical Fish**

   A freshwater aquarium page covering all aspects of the **tropical fish** hobby. ... to Badman's **Tropical Fish**. ... world of aquariology with Badman's **Tropical Fish**. ...

2. **Tropical Fish**

   Notes on a few species and a gallery of photos of African cichlids.

3. The **Tropical** Tank Homepage - **Tropical Fish** and Aquariums

   Info on **tropical fish** and **tropical** aquariums, large **fish** species index with ... Here you will find lots of information on **Tropical Fish** and Aquariums. ...

4. **Tropical Fish** Centre

   Offers a range of aquarium products, advice on choosing species, feeding, and health care, and a discussion board.

5. **Tropical fish** - Wikipedia, the free encyclopedia

   Tropical fish are popular aquarium **fish** , due to their often bright coloration. ... Practical Fishkeeping • **Tropical Fish** Hobbyist • Koi. Aquarium related companies: ...

6. **Tropical Fish** Find

   Home page for **Tropical Fish** Internet Directory ... stores, forums, clubs, **fish** facts, **tropical fish** compatibility and aquarium ...

7. Breeding **tropical fish**

   ... intrested in keeping and/or breeding **Tropical**, Marine, Pond and Coldwater **fish**. ... Breeding **Tropical Fish** ... breeding **tropical**, marine, coldwater & pond **fish**. ...

8. FishLore

   Includes **tropical** freshwater aquarium how-to guides, FAQs, **fish** profiles, articles, and forums.

9. Cathy's **Tropical Fish** Keeping

   Information on setting up and maintaining a successful freshwater aquarium.

10. **Tropical Fish** Place

    Tropical Fish information for your freshwater **fish** tank ... great amount of information about a great hobby, a freshwater **tropical fish** tank. ...

Top 10 documents for “tropical fish”

Clustering Results

- Requirements
- Efficiency
  - must be specific to each query and are based on the top-ranked documents for that query
  - typically based on snippets
- Easy to understand
  - Can be difficult to assign good labels to groups
  - Monothetic vs. polythetic classification

Types of Classification

- Monothetic
  - every member of a class has the property that defines the class
  - typical assumption made by users
  - easy to understand
- Polythetic
  - members of classes share many properties but there is no single defining property
  - most clustering algorithms (e.g. K-means) produce this type of output

Classification Example

$$D_1 = \{a, b, c\}$$
$$D_2 = \{a, d, e\}$$
$$D_3 = \{d, e, f, g\}$$
$$D_4 = \{f, g\}$$

- Possible monothetic classification
  - $\{D_1, D_2\}$ (labeled using *a*) and $\{D_2, D_3\}$ (labeled *e*)
- Possible polythetic classification
  - $\{D_2, D_3, D_4\}$, D_1
  - labels?

Result Clusters

- Simple algorithm
  - group based on words in snippets
- Refinements
  - use phrases
  - use more features
    - whether phrases occurred in titles or snippets
    - length of the phrase
    - collection frequency of the phrase
    - overlap of the resulting clusters,

aquarium (5)	(1, 3, 4, 5, 8)
freshwater (4)	(1, 8, 9, 10)
species (3)	(2, 3, 4)
hobby (3)	(1, 5, 10)
forums (2)	(6, 8)

Faceted Classification

- A set of categories, usually organized into a hierarchy, together with a set of *facets* that describe the important properties associated with the category
- Manually defined
  - potentially less adaptable than dynamic classification
- Easy to understand
  - commonly used in e-commerce

Example Faceted Classification

Books (7,845)
Home & Garden (2,477)
Apparel (236)
Home Improvement (169)
Jewelry & Watches (76)
Sports & Outdoors (71)
Office Products (68)
Toys & Games (62)
Everything Else (44)
Electronics (26)
Baby (25)

DVD (12)
Music (11)
Software (10)
Gourmet Food (6)
Beauty (4)
Automotive (4)
Magazine Subscriptions (3)
Health & Personal Care (3)
Wireless Accessories (2)
Video Games (1)

Categories for “tropical fish”

Example Faceted Classification

Home & Garden
Kitchen & Dining (149)
Furniture & Décor (1,776)
Pet Supplies (368)
Bedding & Bath (51)
Patio & Garden (22)
Art & Craft Supplies (12)
Home Appliances (2)
Vacuums, Cleaning & Storage (107)

Brand
<brand names>
Seller
<vendor names>

Discount
Up to 25% off (563)
25% - 50% off (472)
50% - 70% off (46)
70% off or more (46)

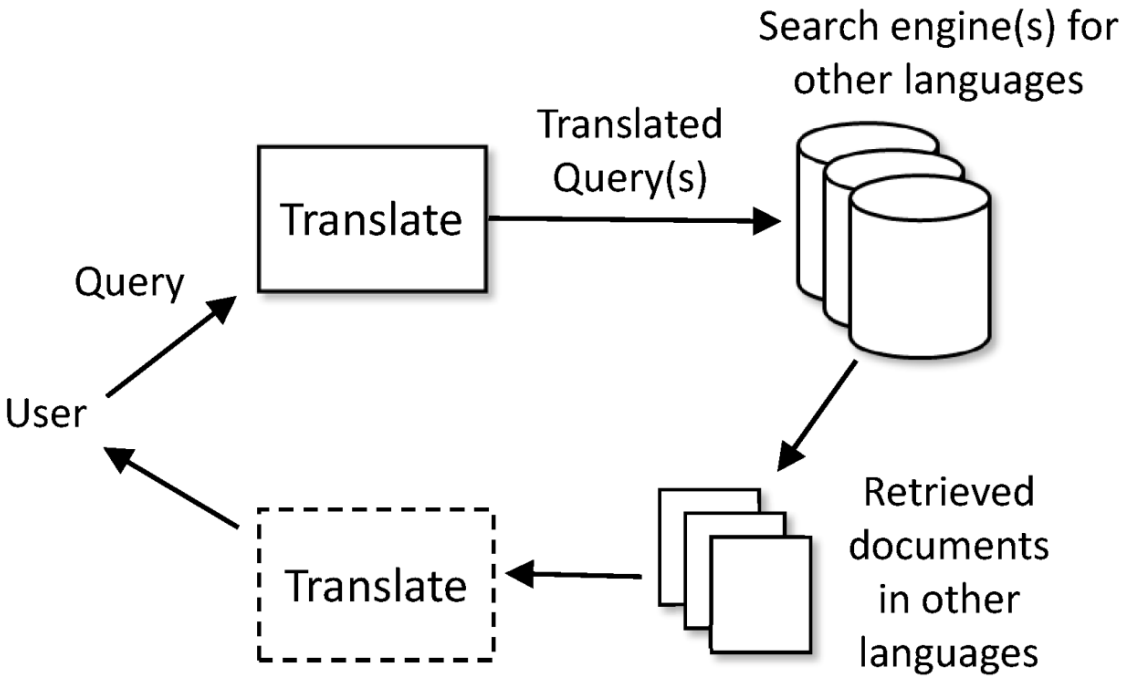
Price
$0-$24 (1,032)
$25-$49 (394)
$50-$99 (797)
$100-$199 (206)
$200-$499 (39)
$500-$999 (9)
$1000-$1999 (5)
$5000-$9999 (7)

Subcategories and facets for “Home & Garden”

Cross-Language Search

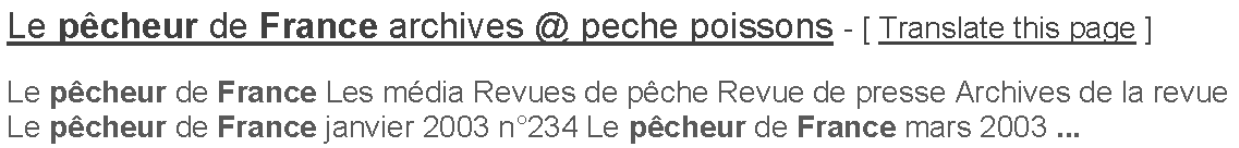
- Query in one language, retrieve documents in multiple other languages
- Involves query translation, and probably document translation
- Query translation can be done using bilingual dictionaries
- Document translation requires more sophisticated *statistical translation* models
  - similar to some retrieval models

Cross-Language Search



Translation

- Web search engines use translation
  - e.g. for query “pecheur france”



  - translation link translates web page
  - uses statistical machine translation models

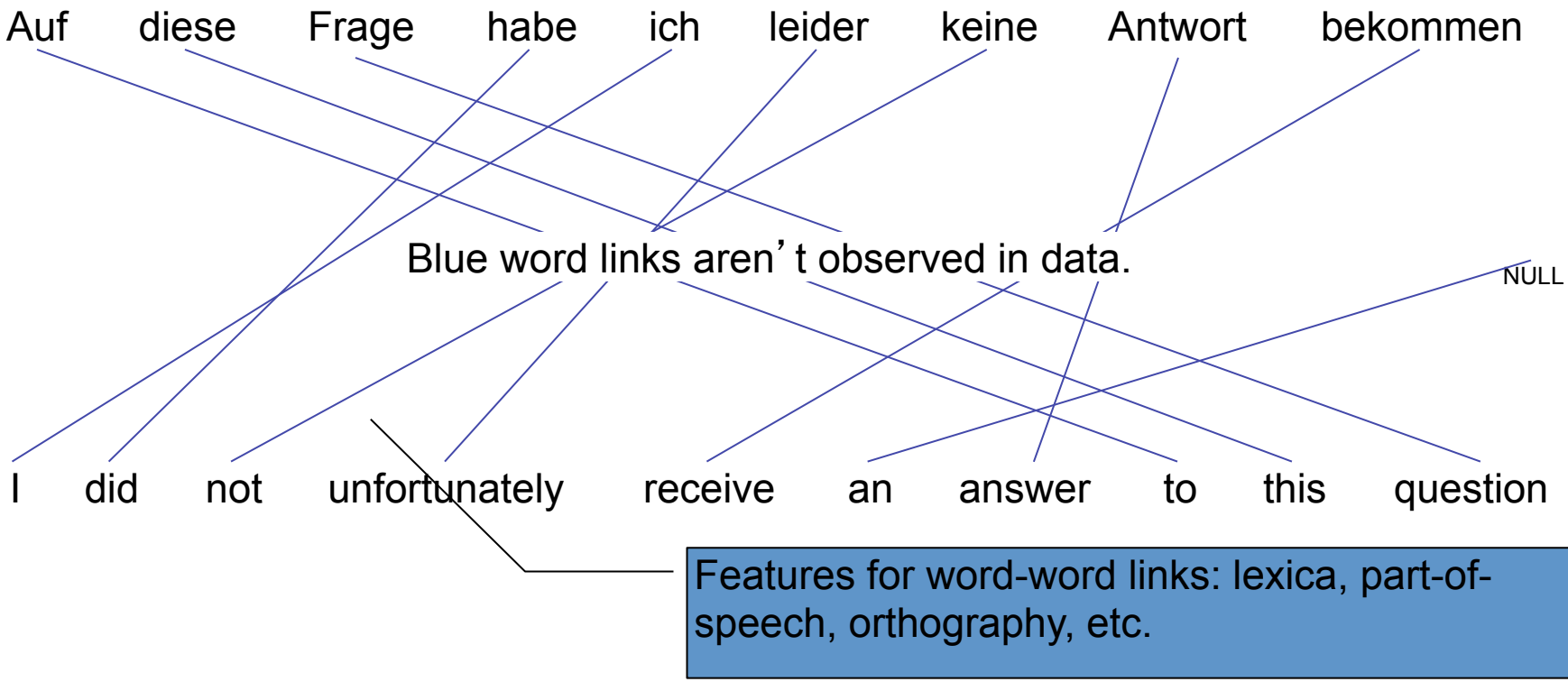
Statistical Translation Models

- Models require *parallel corpora* for training
  - probability estimates based on *aligned* sentences
- Translation of unusual words and phrases is a problem
  - also use *transliteration* techniques
    - e.g., Qathafi, Kaddafi, Qadafi, Gadafi, Gaddafi, Kathafi, Kadhafi, Qadhafi, Qazzafi, Kazafi, Qaddafy, Qadafy, Quadhaffi, Gadhdhafi, al-Qaddafi, Al-Qaddafi

Statistical Translation Models

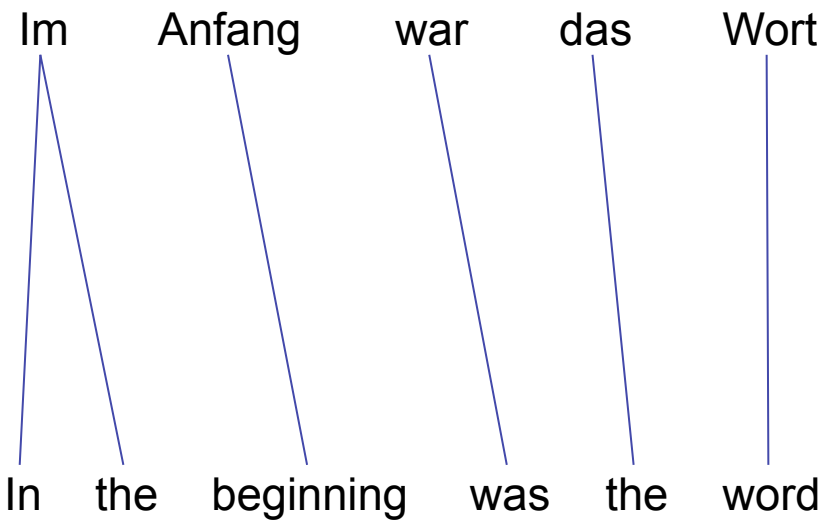
- Translation models
  - "Adequacy"
  - Assign better scores to accurate (and complete) translations
- Language models
  - "Fluency"
  - Assign better scores to natural target language text
- Compare: Error models and language models for spelling correction
  - Warren Weaver: "When I see an article in Russian, I say, 'This is really written in English, but in some strange symbols. I will now proceed to decode.' "

Word Translation Models

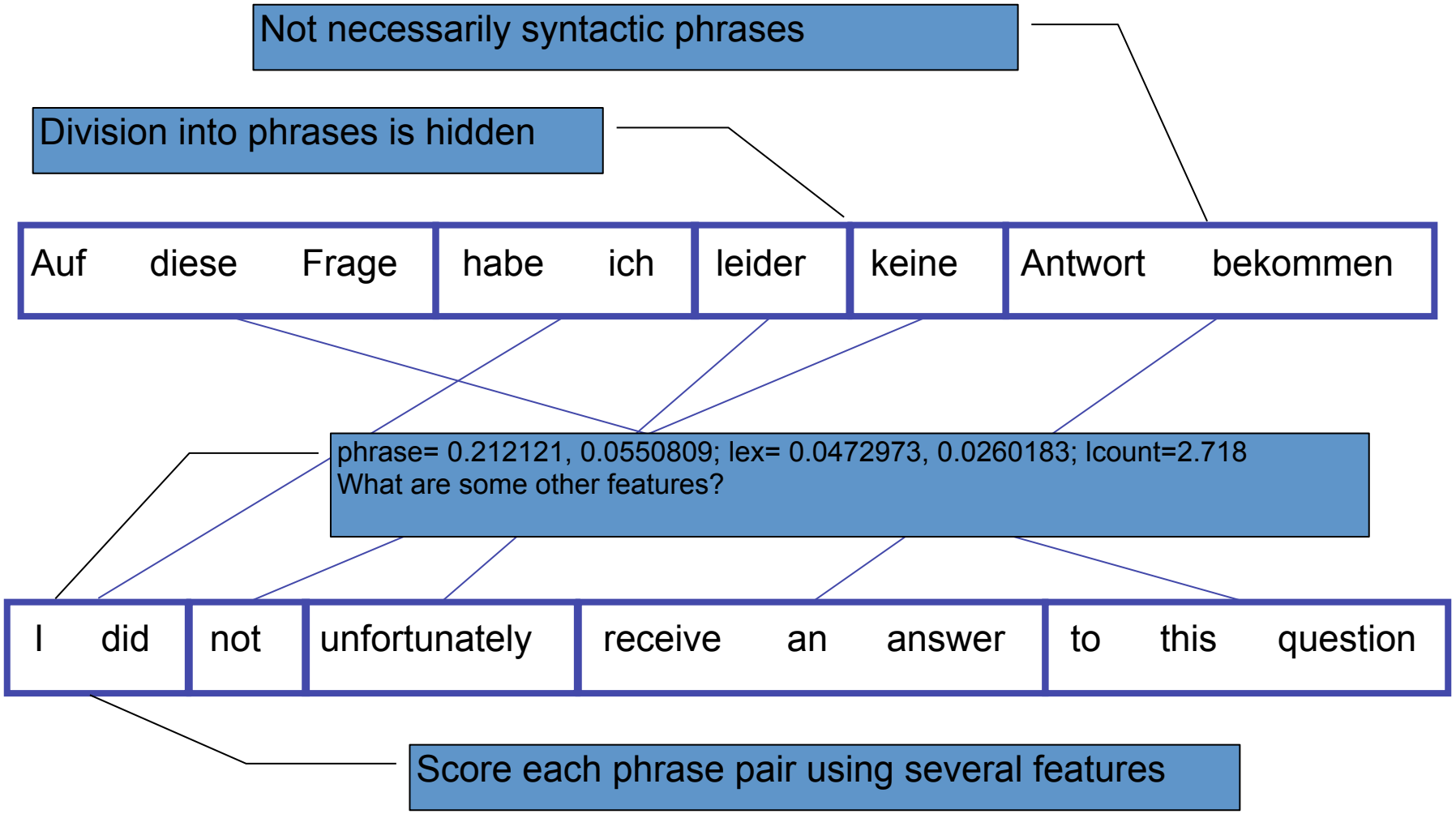


Word Translation Models

- Usually directed: each word in the target generated by one word in the source
- Many-many and null-many links allowed
- Classic IBM models of Brown et al.
- Used now mostly for word alignment, not translation



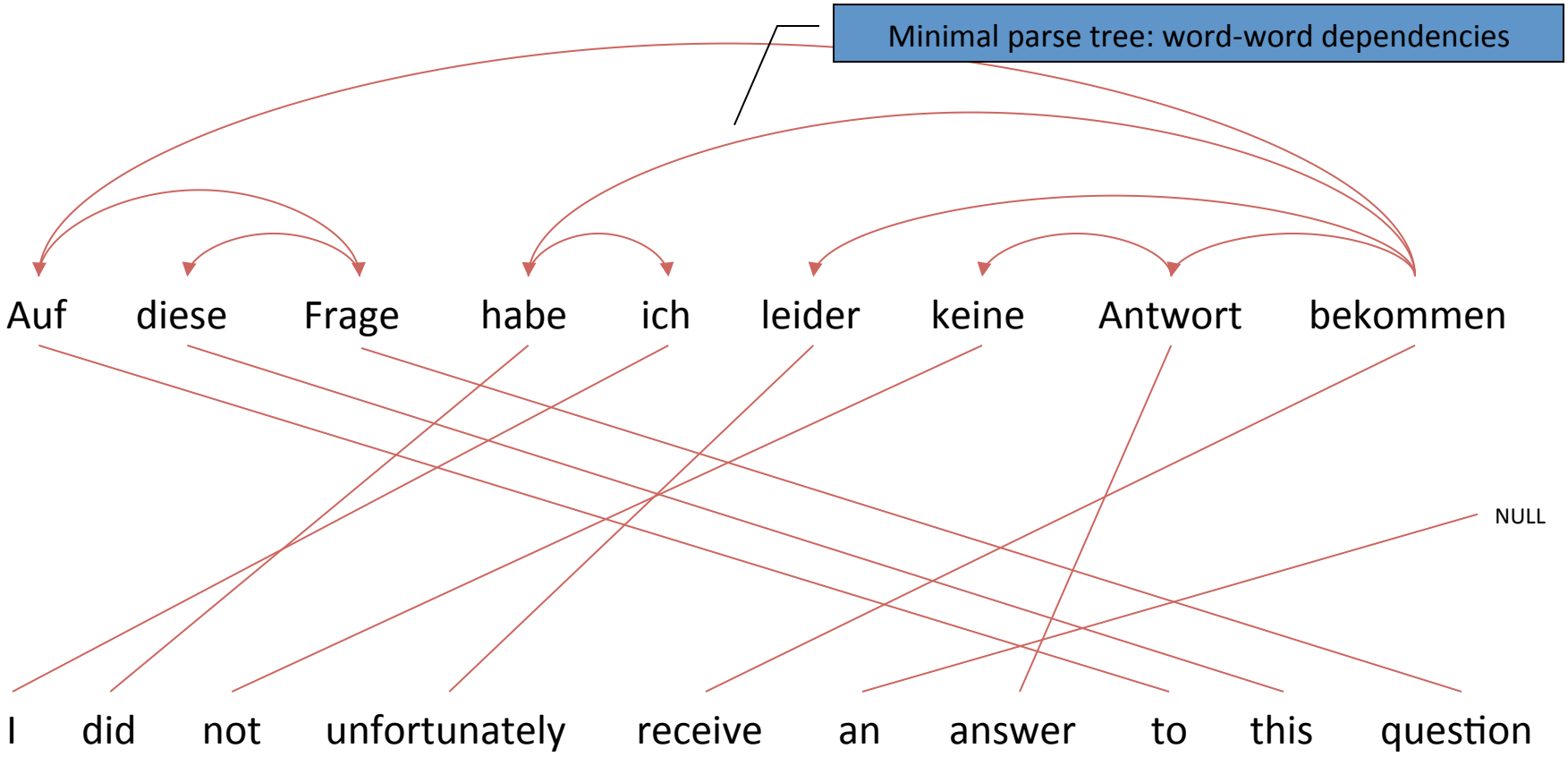
Phrase Translation Models



Phrase Translation Models

- Capture translations in context
  - en Amerique: to America
  - en anglais: in English
- State-of-the-art for several years
- Each source/target phrase pair is scored by several weighted features.
- The weighted sum of model features is the whole translation's score.
- Phrases don't overlap (cf. language models) but have "reordering" features.

Single-Tree Translation Models

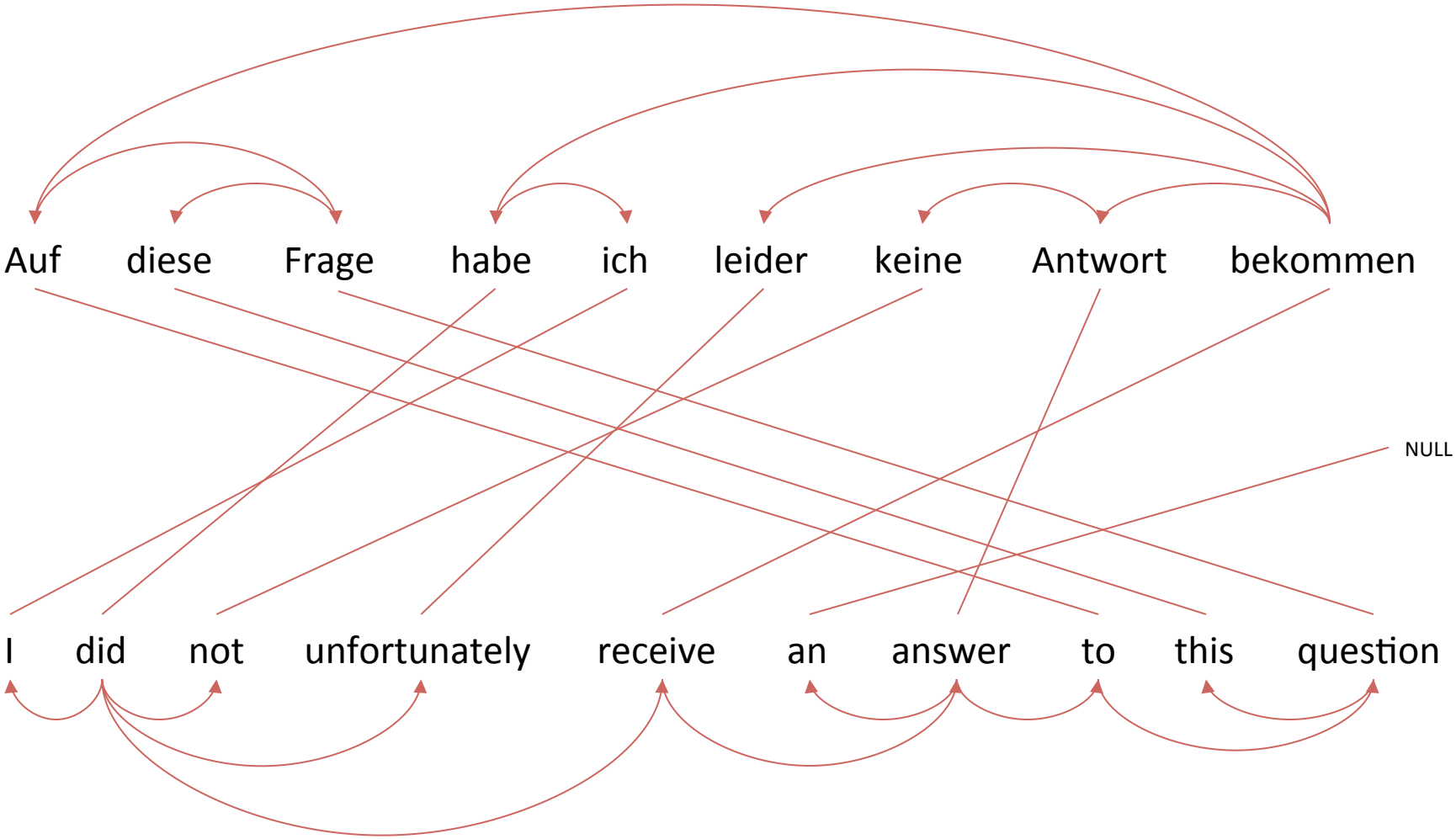


Parse trees with deeper structure have also been used.

Single-Tree Translation Models

- Either source or target has a hidden tree/parse structure
  - Also known as "tree-to-string" or "tree-transducer" models
- The side with the tree generates words/phrases in tree, not string, order.
- Nodes in the tree also generate words/phrases on the other side.
- English side is often parsed, whether it's source or target, since English parsing is more advanced.

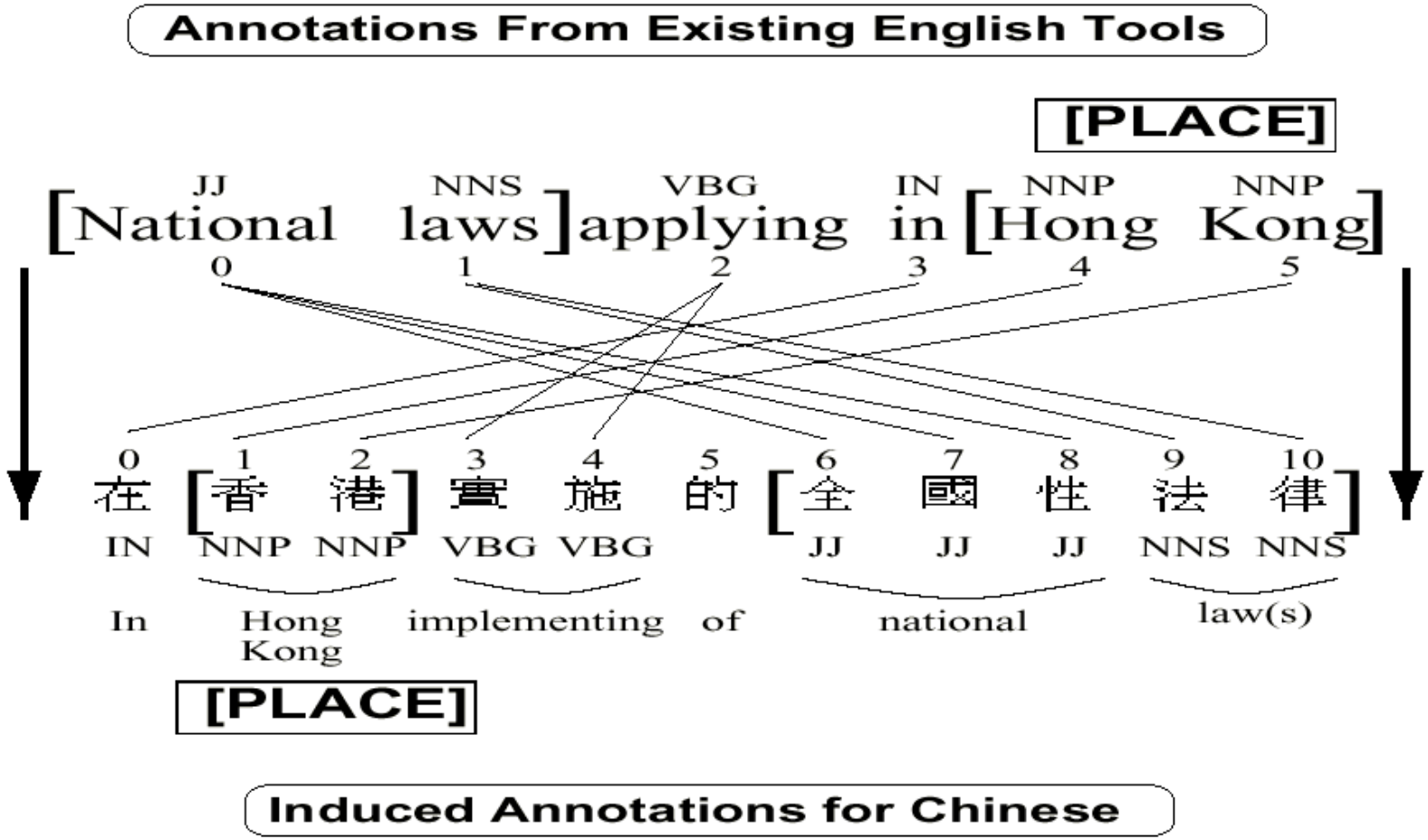
Tree-Tree Translation Models



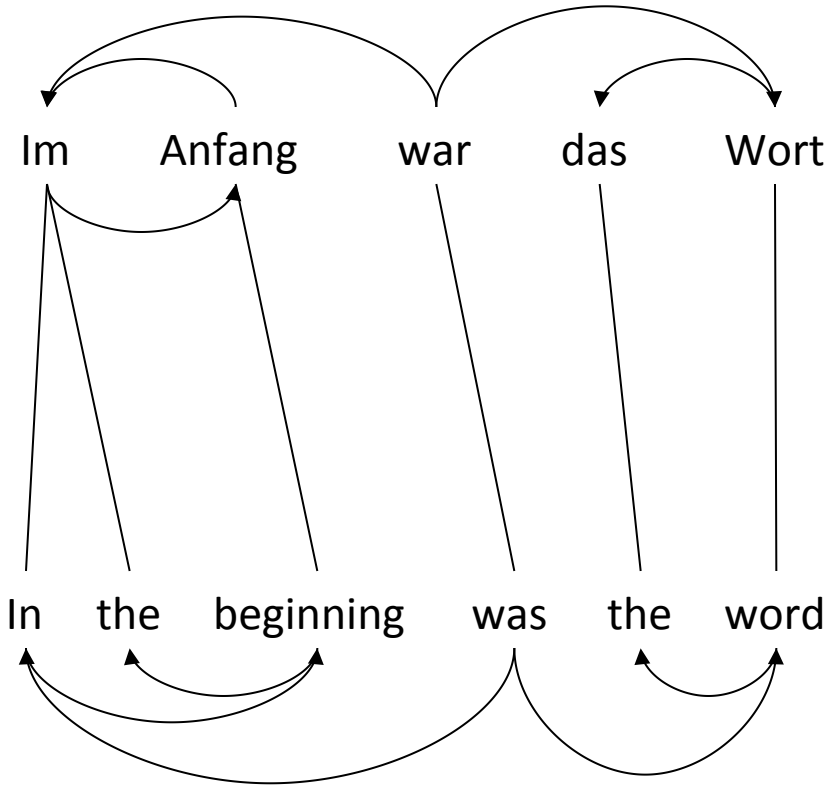
Tree-Tree Translation Models

- Both sides have hidden tree structure
  - Can be represented with a "synchronous" grammar
- Some models assume isomorphic trees, where parent-child relations are preserved; others do not.
- Trees can be fixed in advance by monolingual parsers or induced from data (e.g. Hiero).
- Cheap trees: project from one side to the other

Projecting Hidden Structure



Projection



- Train with bitext
- Parse one side
- Align words
- Project dependencies
- Many to one links?
- Non-projective and circular dependencies?