Computer Processing of Natural Language

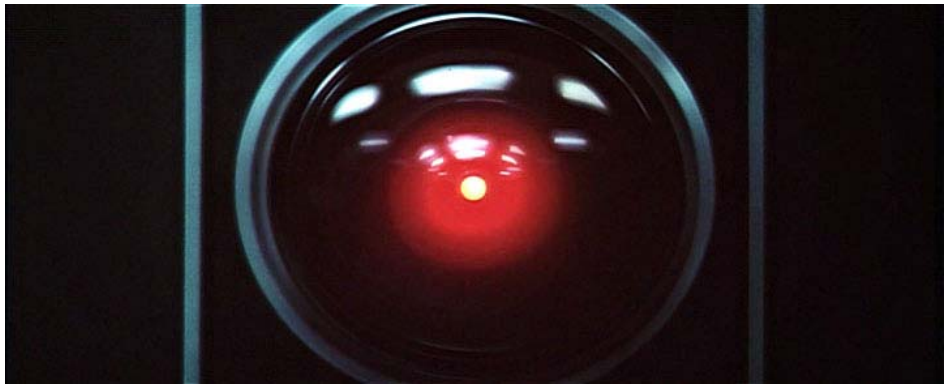
Prof. Hearst

i141

November 26, 2008

We’ve past the year 2001,
but we are not close
to realizing the dream
(or nightmare …)

Dave Bowman: “Open the pod bay doors, HAL”



HAL 9000: “I’m sorry Dave. I’m afraid I can’t do that.”

I know you and Frank were planning to disconnect me,
and I'm afraid that's something I cannot allow to happen.

Why is Computer Processing of Human Language Difficult?

- Computers are not brains
  - There is evidence that much of language understanding is built-in to the human brain
- Computers do not socialize
  - Much of language is about communicating with people
- Key problems:
  - Representation of *meaning*
  - Language only reflects the surface of meaning
  - Language presupposes knowledge about the world
  - Language presupposes communication between people

Piano Practice

by Rilke, translated by Edward Snow

The summer hums. The afternoon fatigues; she breathed her crisp white dress distractedly and put into it that sharply etched etude her impatience for a reality

that could come: tomorrow, this evening-, that perhaps was there, was just kept hidden; and at the window, tall and having everything, she suddenly could feel the pampered park.

With that she broke off; gazed outside, locked her hands together; wished for a long book- and in a burst of anger shoved back the jasmine scent. She found it sickened her.

World Knowledge is subtle

- He arrived at the lecture.
- He chuckled at the lecture.

- He arrived drunk.
- He chuckled drunk.

- He chuckled his way through the lecture.
- ✘ He arrived his way through the lecture.

Words are ambiguous (have multiple meanings)

- I know that.
- I know that block.
- I know that blocks the sun.
- I know that block blocks the sun.

How can a machine understand these differences?

- Get the cat with the gloves.

How can a machine understand these differences?

- Get the sock from the cat with the gloves.
- Get the glove from the cat with the socks.

How can a machine understand these differences?

- Decorate the cake with the frosting.
- Decorate the cake with the kids.
- Throw out the cake with the frosting.
- Throw out the cake with the kids.

Headline Ambiguity

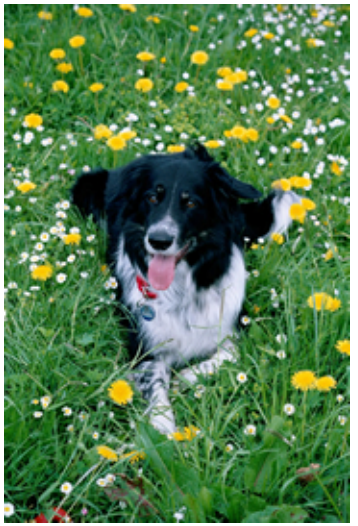
- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Kids Make Nutritious Snacks
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Bush Wins on Budget, but More Lies Ahead
- Hospitals are Sued by 7 Foot Doctors

The Role of Memorization

- Children learn words quickly
  - Around age two they learn about 1 word every 2 hours.
  - (Or 9 words/day)
  - Often only need one exposure to associate meaning with word
    - Can make mistakes, e.g., overgeneralization

      "I goed to the store."
  - Exactly how they do this is still under study
- Adult vocabulary
  - Typical adult: about 60,000 words
  - Literate adults: about twice that.

The Role of Memorization



- Dogs can do word association too!
  - Rico, a border collie in Germany
  - Knows the names of each of 100 toys
  - Can retrieve items called out to him with over 90% accuracy.
  - Can also learn and remember the names of unfamiliar toys after just one encounter, putting him on a par with a three-year-old child.

http://www.nature.com/news/2004/040607/pf/040607-8_pf.html

But there is too much to memorize!

establish

establishment

> the church of England as the official state church.

disestablishment

antidisestablishment

antidisestablishmentarian

antidisestablishmentarianism

> is a political philosophy that is opposed to the separation of church and state.

Rules and Memorization

- Current thinking in psycholinguistics is that we use a combination of rules and memorization
  - However, this is very controversial
- Mechanism:
  - If there is an applicable rule, apply it
  - However, if there is a memorized version, that takes precedence. (Important for irregular words.)
    - Artists paint "still lifes"
      - Not "still lives"
    - Past tense of
      - think → thought
      - blink → blinked
- This is a simplification; for more on this, see Pinker's "Words and Rules" and "The Language Instinct".

Language subtleties

- Adjective order and placement
  - A big black dog
  - A big black scary dog
  - A big scary dog
  - A scary big dog
  - ✘ A black big dog

- Antonyms
  - Which sizes go together?
    - Big and little
    - Big and small
    - Large and small
    - ✘ Large and little

Representation of Meaning

- I know that block blocks the sun.
  - How do we represent the meanings of “block”?
  - How do we represent “I know”?
  - How does that differ from “I know that.”?
  - Who is “I”?
  - How do we indicate that we are talking about earth’s sun vs. some other planet’s sun?
  - When did this take place? What if I move the block? What if I move my viewpoint? How do we represent this?

How to tackle these problems?

- First attempt: write all the rules down.
  - Rules for syntactic structure.
  - Rules for meanings of words.
  - Rules for how to combine the meanings.

Green Eggs and Ham, Dr. Seuss

I am Sam	Subject Verb Object
I am Sam	Subject Verb Object
Sam I am	Object, Subject Verb
That Sam-I-am!	Demonstrative Proper-Noun
That Sam-I-am!	Noun Do Modal Verb
I do not like that Sam-I-am!	Demonstrative Proper-Noun
Do you like green eggs and ham?	
I do not like them,	
Sam-I-am.	
I do not like green eggs and ham.	

Green Eggs and Ham, Dr. Seuss

I am Sam
I am Sam
Sam I am

That Sam-I-am!
That Sam-I-am!
I do not like that Sam-I-am!

Do you like green eggs and ham?

I do not like them,
Sam-I-am.
I do not like green eggs and ham.

Rule: declaration of self's name
Rule: repeating declaration indicates Emphasis but no change in meaning.

Rule: stating someone's name
In a declarative suggests … anger? Admiration? …?
Rule: first person stating not liking Indicates negative feelings towards Other person.

“Closed Domain” Question Answering Systems

- **One example: LUNAR (Woods & Kaplan 1977)**
- **Answered questions about moon rocks and soil gathered by the Apollo 11 mission.**
  - Parse English questions into a database query
  - Heuristics about how to convert language into meaning
  - **Question:**
    - **Do any samples have greater than 13 percent aluminum?**
  - **Database query**
    - **(TEST (FOR SOME X1 / (SEQ SAMPLES):**
    - **T;**
    - **(CONTAIN X1**
    - **(NPR* X2 / ‘AL203)**
    - **(GREATERTHAN 13 PCT)))**
  - **Answer:**
    - Yes.

How to tackle these problems?

- First attempt: write all the rules down.
  - This didn't work.
  - The field was stuck for quite some time.
- A new approach started around 1990
  - Well, not really new, but the first time around, in the 50's, they didn't have the text, disk space, or GHz
- Main idea: combine memorizing and rules
- How to do it:
  - Get large text collections (corpora)
  - Compute statistics over the words in those collections
- Surprisingly effective
  - Even better now with the Web

Example Problem

- Grammar checker example:

  Which word to use?

  <principal> <principle>

- Solution: look at which words surround each use:
  - **I am in my third year as the principal of Anamosa High School.**
  - **School-principal transfers caused some upset.**
  - **This is a simple formulation of the quantum mechanical uncertainty principle.**
  - **Power without principle is barren, but principle without power is futile. (Tony Blair)**

Using Very, Very Large Corpora

- Keep track of which words are the neighbors of each spelling in well-edited text, e.g.:
  - Principal: “high school”
  - Principle: “rule”
- At grammar-check time, choose the spelling best predicted by the surrounding words.
- Surprising results:
  - Log-linear improvement even to a billion words!
  - Getting more data is better than fine-tuning algorithms!

The Effects of LARGE Datasets

- From Banko & Brill '01

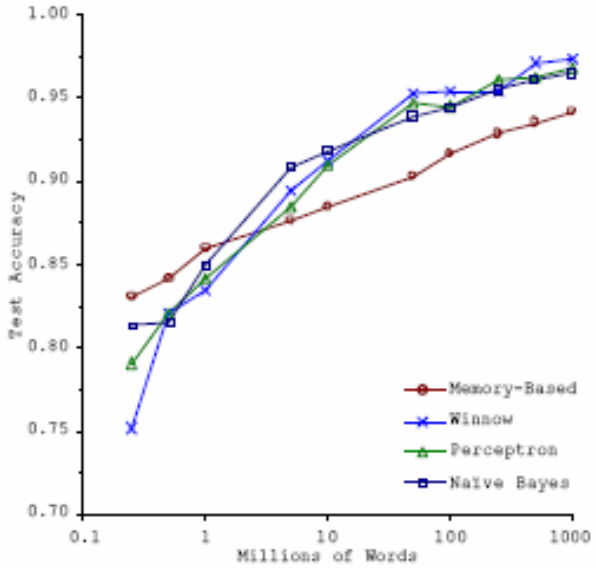


Figure 1. Learning Curves for Confusion Set Disambiguation

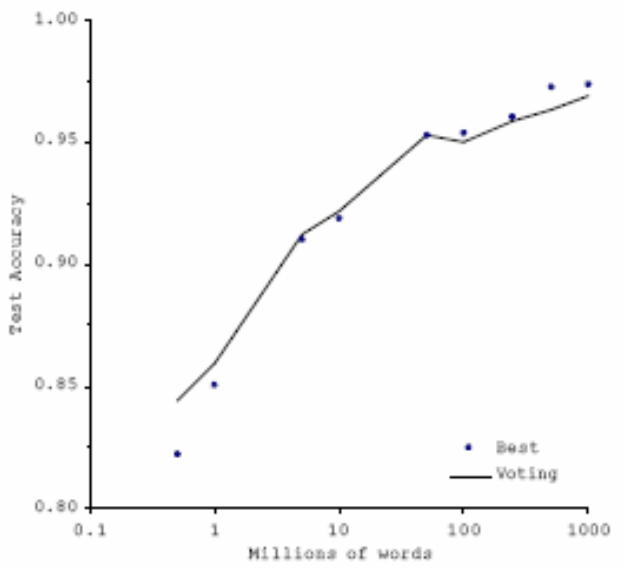


Figure 3. Voting Among Classifiers

Real-World Applications of NLP

- Spelling Suggestions/Corrections
- Grammar Checking
- Synonym Generation
- Information Extraction
- Text Categorization
- Automated Customer Service
- Speech Recognition (limited)
- Machine Translation
- In the (near?) future:
  - Question Answering
  - Improving Web Search Engine results
  - Automated Metadata Assignment
  - Online Dialogs

Automatic Help Desk Translation at Microsoft

Esta buscando: estilo viñetas powerpoint

Resultados de la búsqueda

Mostrar resultados para:

- PowerPoint
- PowerPoint 2002
- PowerPoint 2003
- PowerPoint 2000
- Windows NT
- PowerPoint 2001

Resultados 1-20 de 200+ Siguiente > Mostrar todos

- Recibe un mensaje de error al intentar abrir una presentación en PowerPoint 2003 o PowerPoint 2002

  (820703) - Describe un error de error abierto que recibe al intentar abrir una presentación de PowerPoint 2003 o PowerPoint 2002. Puede ser capaz de abrir la presentación en una versión anterior de PowerPoint para funcionar de PowerPoint alrededor de este problema.
  http://support.microsoft.com/kb/820703/es

- Mensaje de error a que ve una presentación de PowerPoint 2003 o PowerPoint 2002 ""Cargar Ser Poder de Hlink.dll o "Hlink.dll de Cargar a Fallar de PowerPoint"

  (813726) - Al ver una presentación de PowerPoint 2003 de Microsoft Office o Microsoft PowerPoint 2002, uno o ambos mensaje de error siguientes pueden aparecer : no se puede cargar Microsoft PowerPoint "hlink.dll".
  http://support.microsoft.com/kb/813726/es

- PPT2000 mensaje de error:" PowerPoint Viewer no puede leer

  (226769) - Al ver una presentación de PowerPoint 2000 desempaquetado que utiliza el visor 97 de Microsoft PowerPoint, el mensaje de error siguiente puede aparecer: PowerPoint Viewer no puede leer ruta de acceso C:\ \ .ppt de nombre de archivo
  http://support.microsoft.com/kb/226769/es

- PPT7: Importar Freelance 96 Presentations a PowerPoint

  (161532) - La versión 7.0 Microsoft PowerPoint for Windows 95 no incluye un convertidor para Lotus Freelance Gráficos 96 para 2.x de Windows 95 o Lotus Freelance para archivos de Windows.
  http://support.microsoft.com/kb/161532/es

- Recibe "no se pueden activar Algunos controles de esta presentación" mensaje de error cuando utiliza PowerPoint 97 para abrir una presentación de PowerPoint 2003

  (813720) - Explica que recibe un mensaje de error al intentar abrir una presentación de PowerPoint 2003 en PowerPoint 97. Requiere que abre la presentación en PowerPoint 2002 o PowerPoint 2003 para solucionar este problema.
  http://support.microsoft.com/kb/813720/es

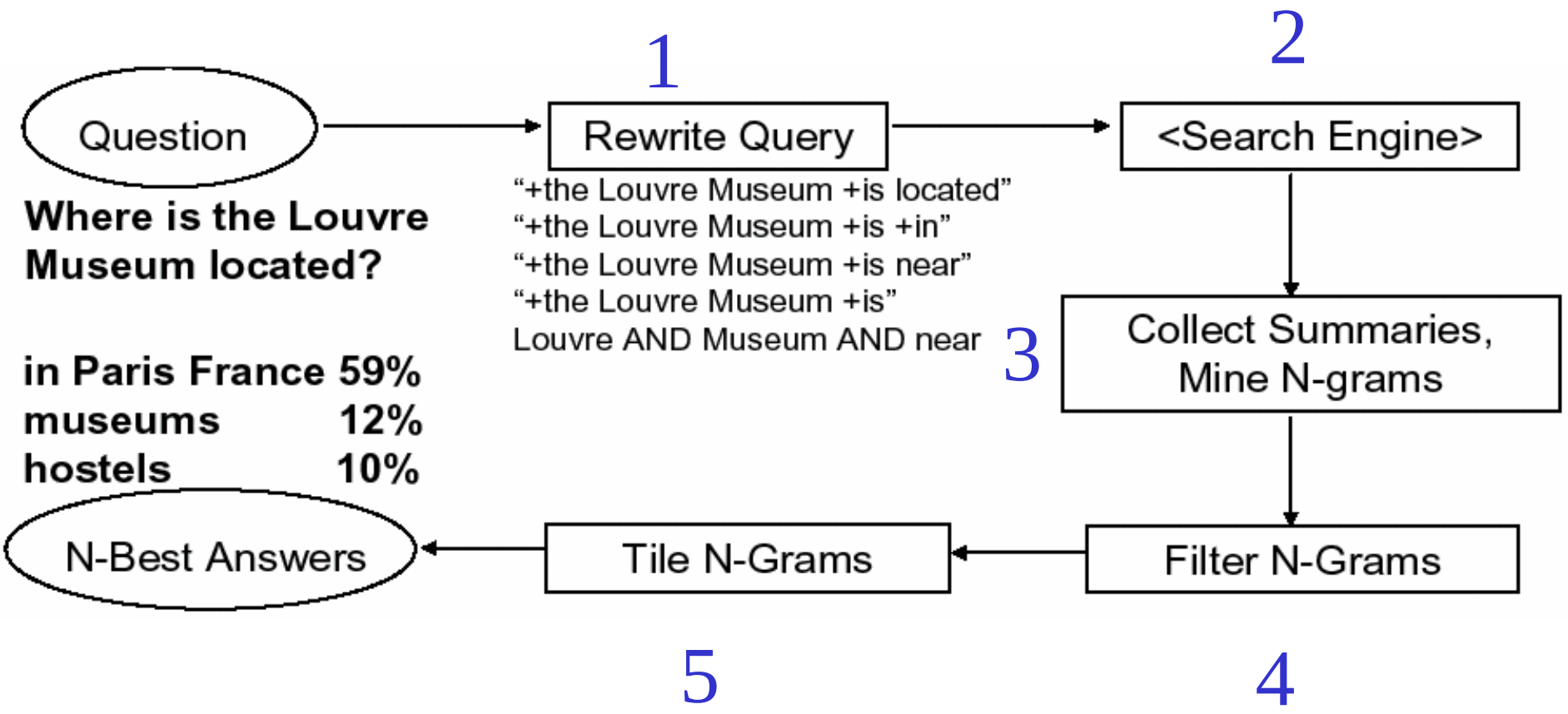
Transferring data from support.microsoft.com...

Synonym Generation

Application to Question Answering

- Goal: make the simplest possible QA system by exploiting the redundancy in the web
  - Use this as a baseline against which to compare more elaborate systems.
  - The next slides based on:
    - Web Question Answering: Is More Always Better? Dumais, Banko, Brill, Lin, Ng, SIGIR'02
    - An Analysis of the AskMSR Question-Answering System, Brill, Dumais, and Banko, EMNLP'02.

AskMSR System Architecture



Step 1: Rewrite the questions

- Intuition: The user's question is often syntactically quite close to sentences that contain the answer.
  - Where is the Louvre Museum located?
  - The Louvre Museum is located in ***Paris***
  - Who created the character of Scrooge?
  - ***Charles Dickens*** created the character of Scrooge.

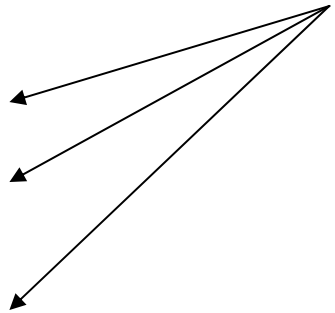
Query rewriting

Classify question into seven categories

- **Who** is/was/are/were…?
- **When** is/did/will/are/were …?
- **Where** is/are/were …?

a. Hand-crafted category-specific transformation rules
e.g.: For *where* questions, move 'is' to all possible locations
Look to the right of the query terms for the answer.

"Where is the Louvre Museum located?"
→ "is the Louvre Museum located"
→ "the is Louvre Museum located"
→ "the Louvre is Museum located"
→ "the Louvre Museum is located"
→ "the Louvre Museum located is"

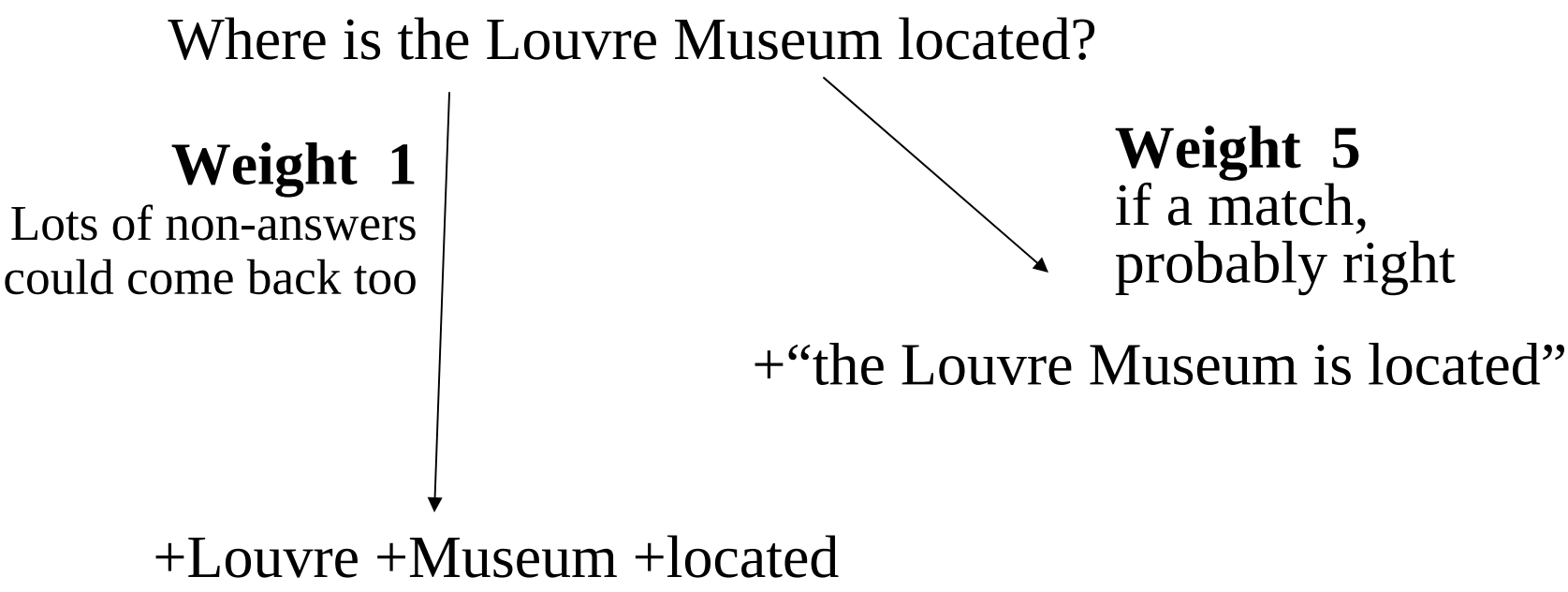


Nonsense, but ok. It's only a few more queries to the search engine.

b. Expected answer "Datatype" (eg, Date, Person, Location, …)
When was the French Revolution? → DATE

Query Rewriting - weighting

Some query rewrites are more reliable than others.



Step 2: Query search engine

- Send all rewrites to a Web search engine
- Retrieve top N answers (100-200)
- For speed, rely just on search engine's "snippets", not the full text of the actual document

Definition: n-gram

- Just means we have N adjacent text string
- Bigram: two adjacent words (big cat)
- Trigram: three adjacent words (big black cat)
- N-gram: not specifying how many adjacent words; leave it loose as a variable.

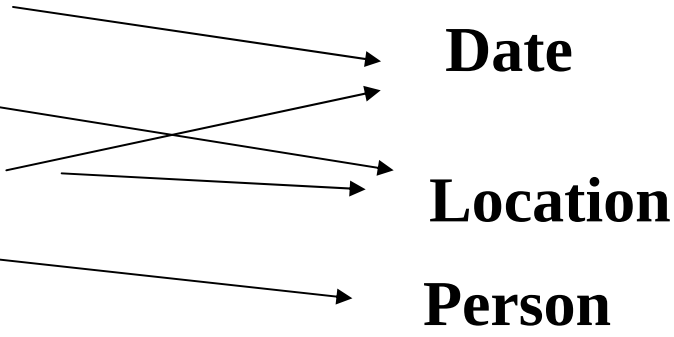
Step 3: Gathering N-Grams

- Enumerate all N-grams (N=1,2,3) in all retrieved snippets
- Weight of an n-gram: occurrence count, each weighted by "reliability" (weight) of rewrite rule that fetched the document
  - Example: "Who created the character of Scrooge?"

Dickens	117
Christmas Carol	78
Charles Dickens	75
Disney	72
Carl Banks	54
A Christmas	41
Christmas Carol	45
Uncle	31

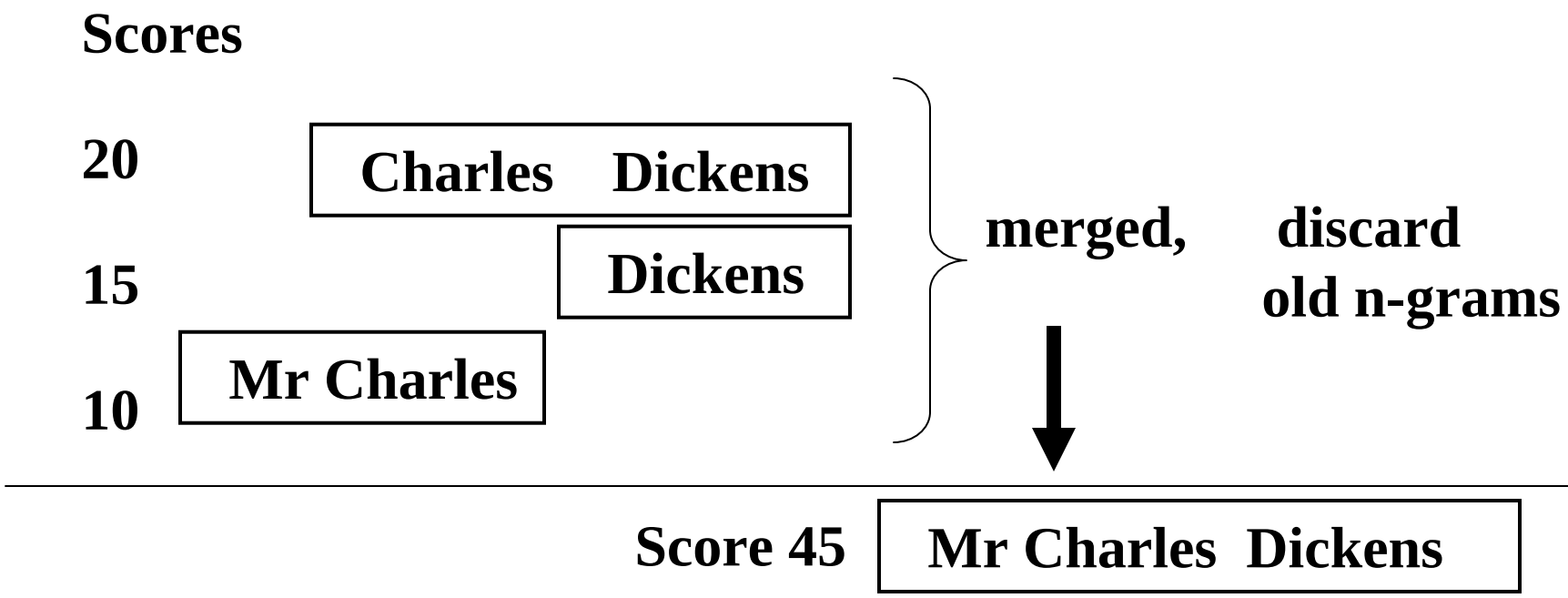
Step 4: Filtering N-Grams

- Each question type is associated with one or more "data-type filters" = regular expression
- When...
- Where...
- What ...
- Who ...



- Boost score of n-grams that match a pattern
- Lower score of n-grams that don't match a pattern

Step 5: Tiling the Answers



Issues

- Works best/only for “Trivial Pursuit”-style fact-based questions
- Limited/brittle repertoire of
  - question categories
  - answer data types/filters
  - query rewriting rules

Summary

- Natural language processing is difficult!
- However, we've made progress over 40 years of research on subproblems
  - Recognizing short spoken sequences
  - Passable machine translation in some cases
  - Getting better at simple question answering!
- What does the future hold?