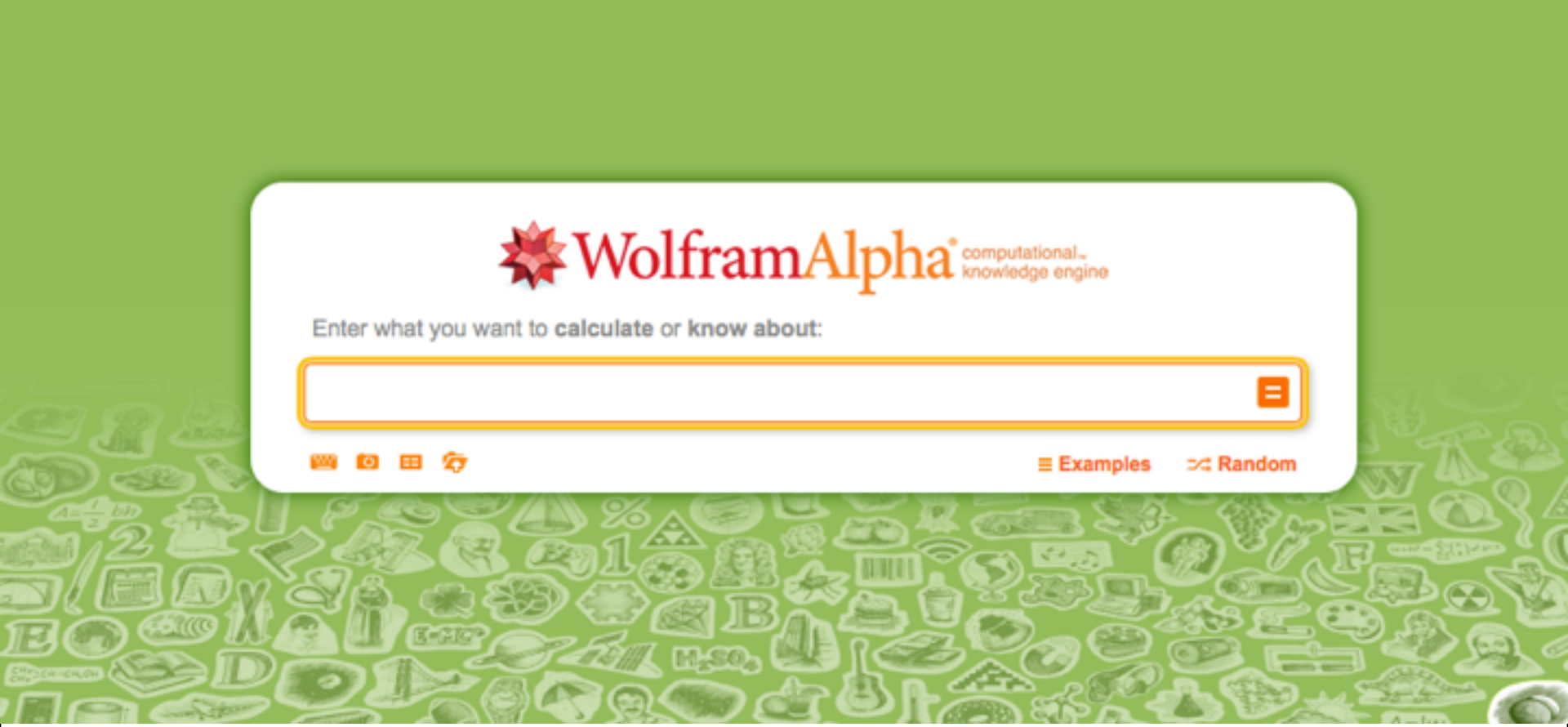
Grupo de Pesquisa em
Inteligência e Imagens

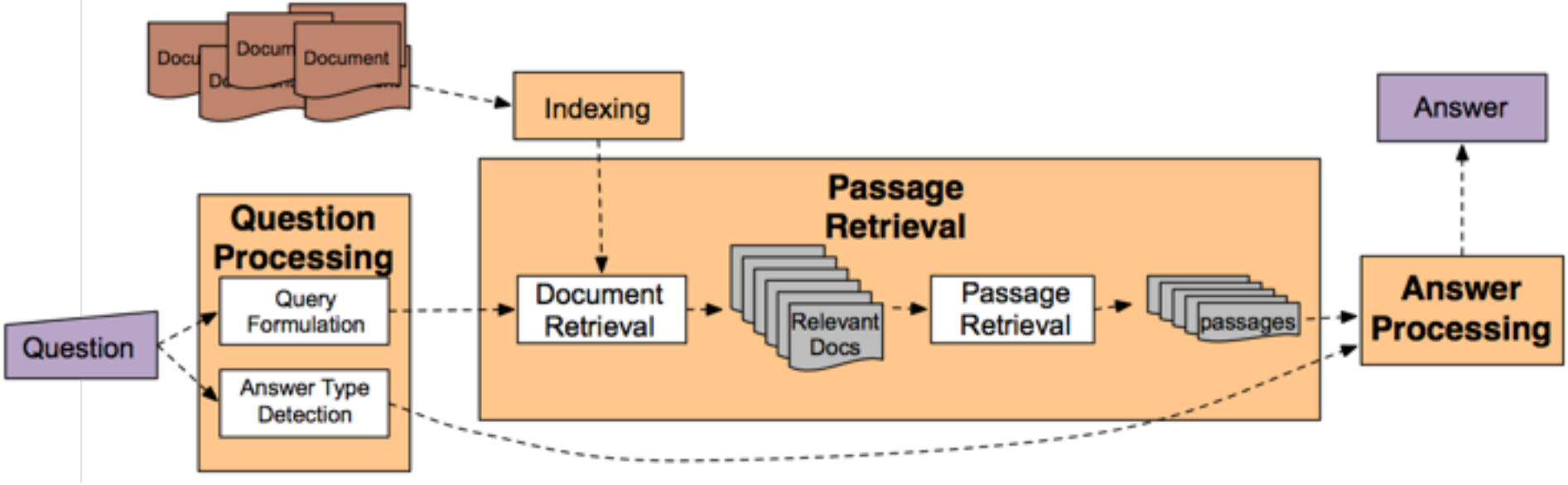
Sistemas de *Question-Answering (QA)*


WolframAlpha computational knowledge engine
Enter what you want to calculate or know about:
Examples
Random

Abordagens para QA

- Baseadas em Recuperação de Informação (IR)
  - TREC; Google
- Baseadas em Conhecimento (KB)
  - Apple Siri; Wolfram Alpha
- Híbridas
  - IBM Watson

QA via IR


Document
Indexing
Answer
Question Processing
Passage Retrieval
Question
Query Formulation
Answer Type Detection
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing

QA via KB

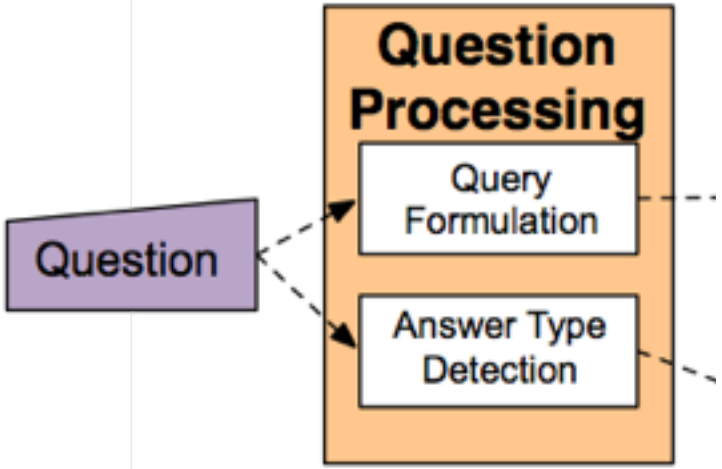
- Constrói representação semântica da query
  - Datas, localidades, entidades, quantidades, etc.
- Representação semântica —> query em dados estruturados
  - Bases de dados geoespaciais
  - Ontologias (Wikipedia, dbPedia, WordNet, etc)
  - Bases de dados científicas

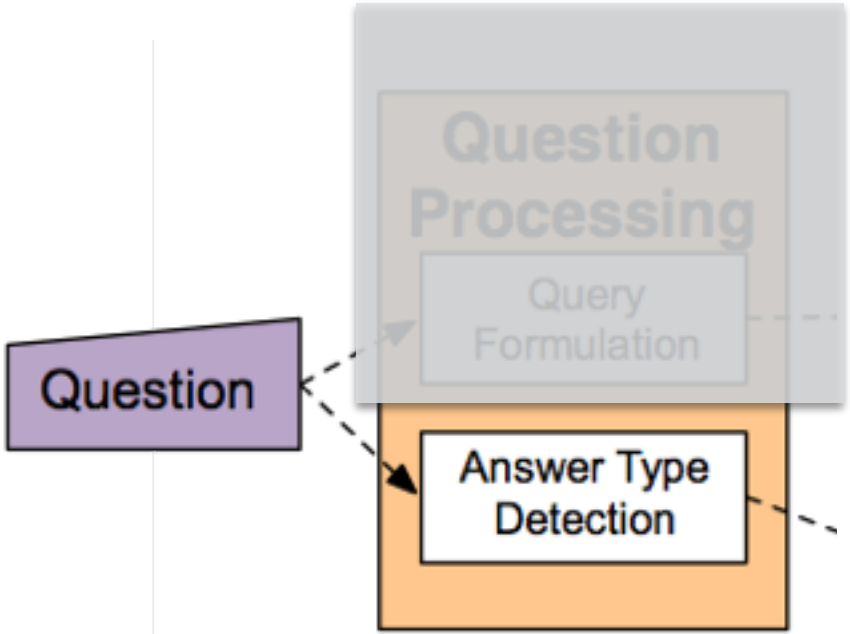
QA híbrida

- Constrói representação semântica da query
- Gera respostas candidatas via IR
- Pontua cada candidata usando fontes de conhecimento mais ricas

Grupo de Pesquisa em
Inteligência e Imagens

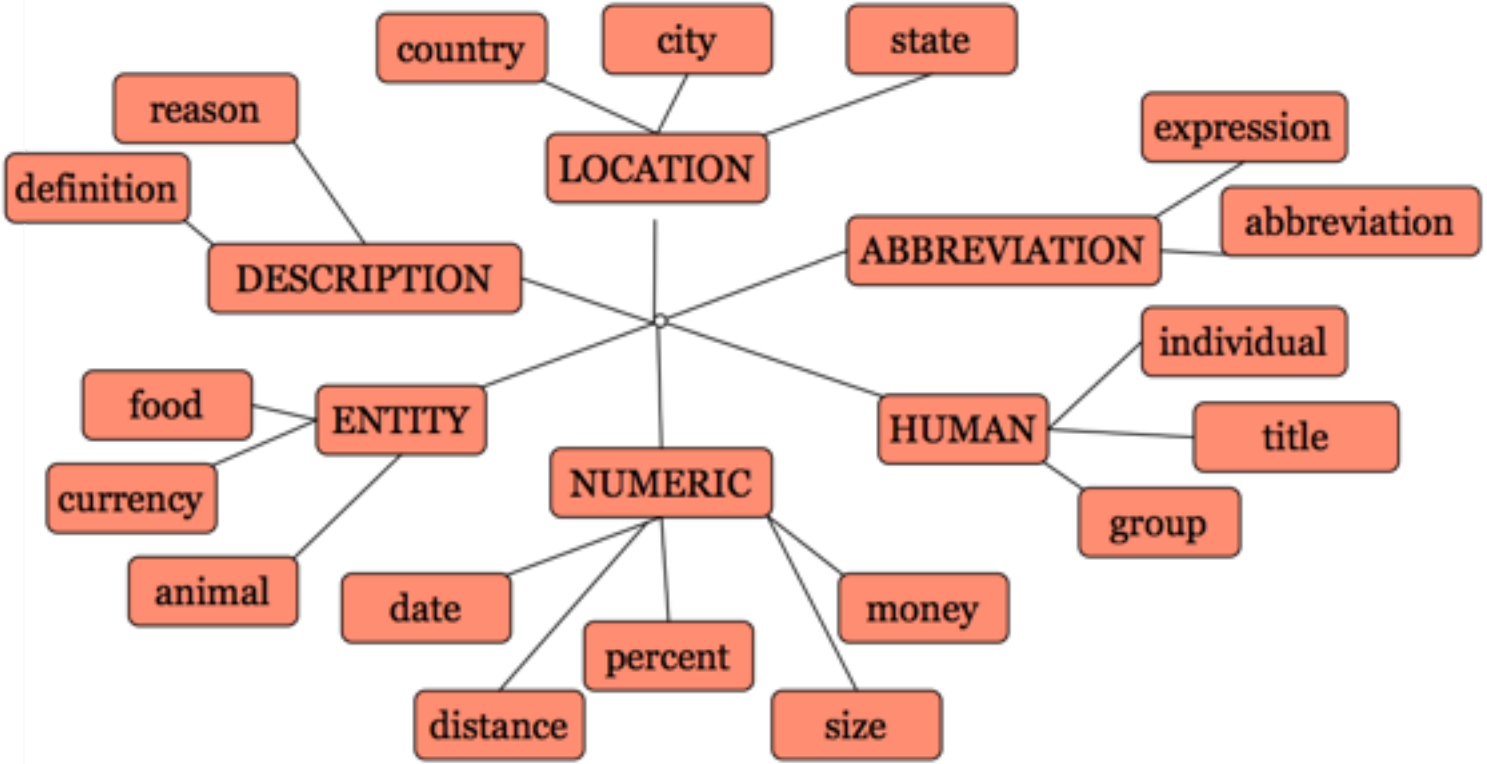
Tipos de respostas e formulação da query




Question Processing
Query Formulation
Question
Answer Type Detection

Taxonomia do tipo de resposta

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02

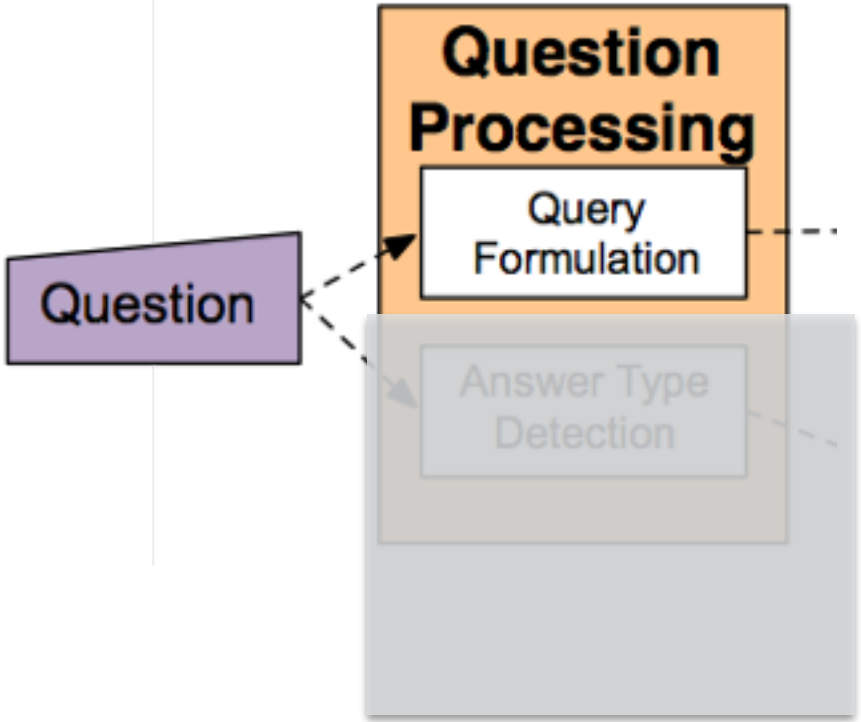


Detecção do tipo de resposta

- Regras definidas manualmente
  - Uso de Expressões Regulares
    - Who {islwaslarelwere} PERSON
    - PERSON (YEAR – YEAR)
  - Palavra central da 1a frase nominal depois do pronome relativo (Que? Qual? Quem? Como? Quando? Onde?)
    - Quem é o autor de Dracula?
    - Qual é a flor do estado da California?

Detecção do tipo de resposta

- Aprendizagem de Máquina
  - Definir taxonomia de tipos de questões
  - Rotular dados de treinamento para cada tipo de questão
  - Treinar classificador para cada tipo de questão usando features:
    - Palavras e frases de questionamento
    - POS tags
    - Palavras centrais (headwords)
    - Entidades nomeadas
    - Palavras relacionadas semanticamente


Question Processing
Query Formulation
Question
Answer Type Detection

Algoritmo para seleção de palavras-chave

Dan Moldovan, Sanda Harabagiu, Marius Paca, Rada Mihalcea, Richard Goodrum, Roxana Girju and Vasile Rus. 1999. Proceedings of TREC-8.

1. *Select all non-stop words in quotations*
2. *Select all NNP words in recognized named entities*
3. *Select all complex nominals with their adjectival modifiers*
4. *Select all other complex nominals*
5. *Select all nouns with their adjectival modifiers*
6. *Select all other nouns*
7. *Select all verbs*
8. *Select all adverbs*
9. *Select the QFW word (skipped in all previous steps)*
10. *Select all other words*

1. Select all non-stop words in quotations

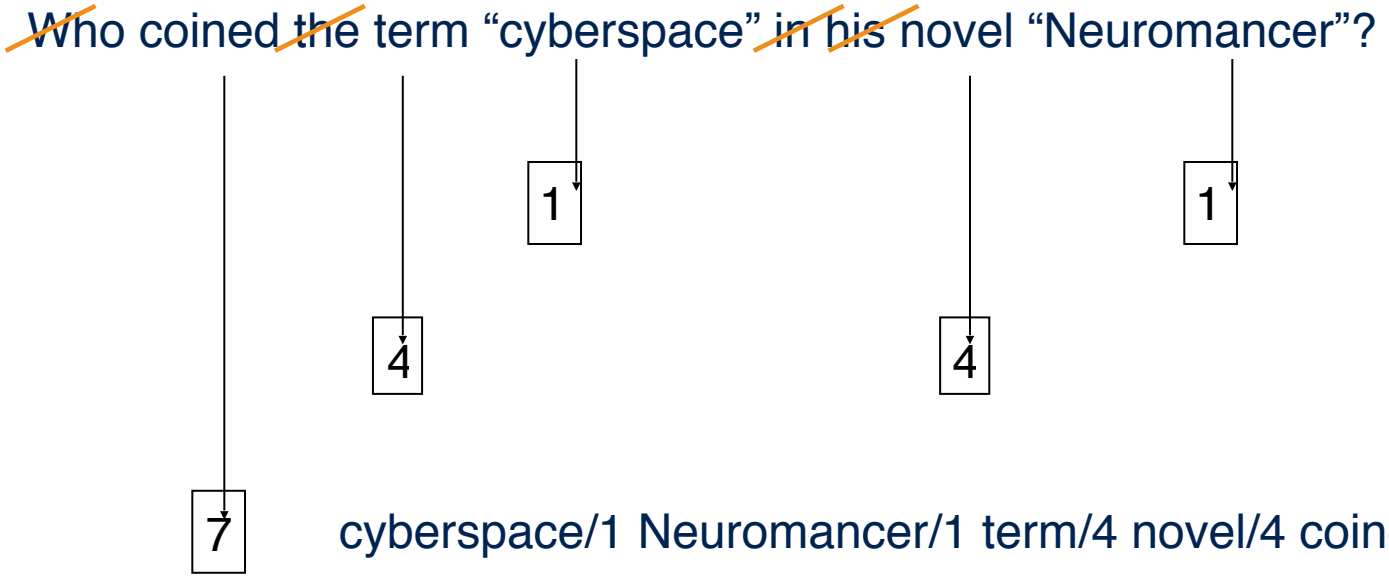
…

4. Select all other complex nominals

…

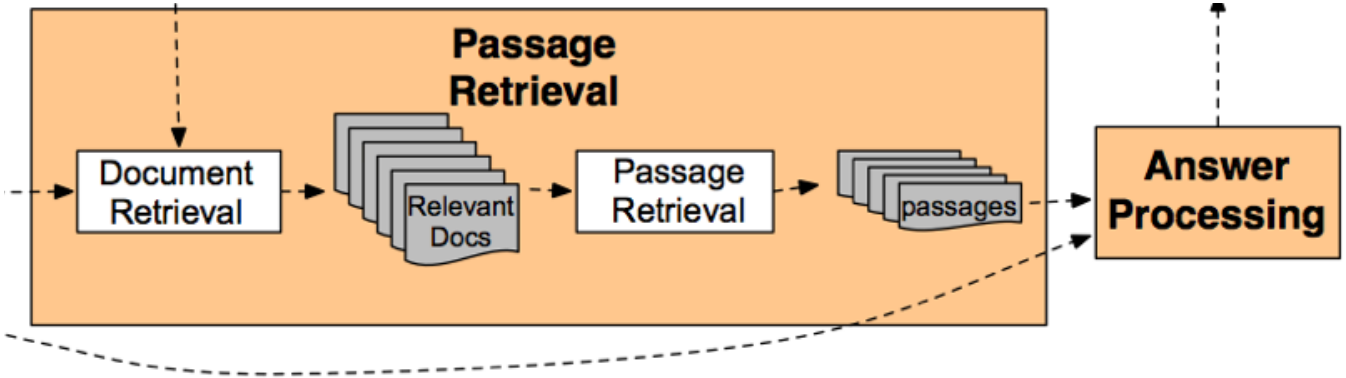
7. Select all verbs

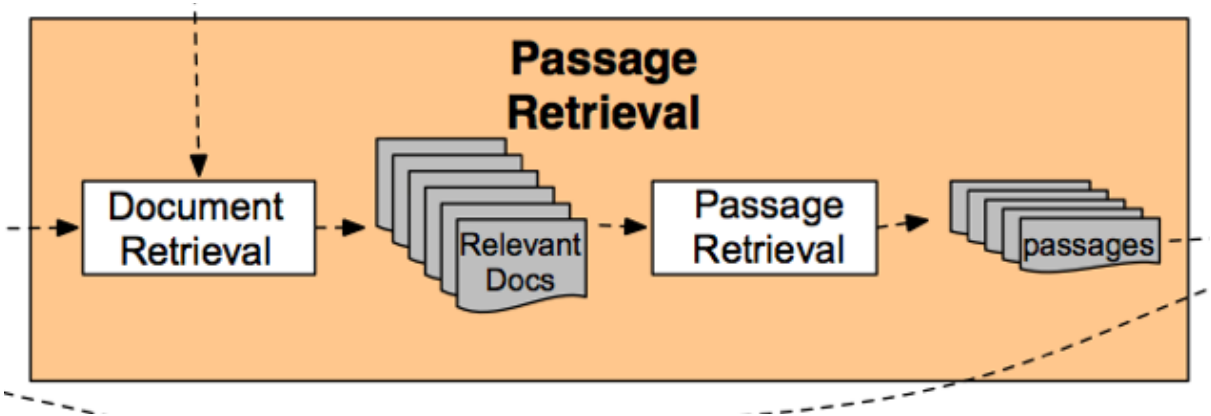
…



Grupo de Pesquisa em
Inteligência e Imagens

Recuperação do trecho
e Extração da resposta

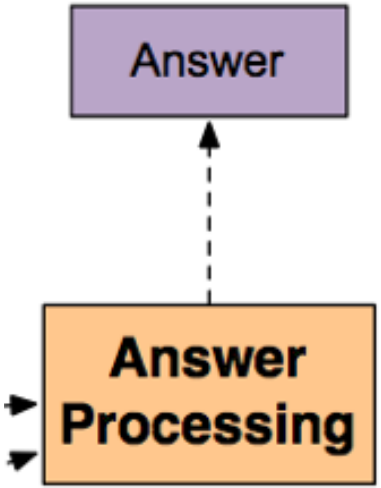




- Passo 1: IR recupera documentos a partir da query
- Passo 2: Segmenta documentos em trechos (ex. parágrafos)
- Passo 3: Ranqueamento do trecho
  - utiliza tipo da resposta para re-ranquear o trecho

Features para ranqueamento do trecho

- # de Entidades Nomeadas do tipo certo no trecho
- # de palavras da query existentes no trecho
- Proximidade das palavras-chave da query para qualquer outra no trecho
- Ranqueamento do documento contendo o trecho


Answer
Answer Processing

Extração da resposta

- Executar um rotulador de entidades nomeadas para tipos de resposta nos trechos
  - Cada tipo de resposta requer um rotulador que a detecte
  - Se o tipo é CITY, o rotulador deve rotular CITY
    - pode ser via RegEx
- Retorna a string com o tipo correto
  - Quem é o Primeiro Ministro da Índia? **(PERSON)**

Manmohan Singh, Primeiro Ministro da Índia, havia contado a líderes da esquerda que o acordo não seria renegociado.

  - Quão alto é o Monte Everest? **(LENGTH)**

A altura oficial do monte Everest é **29035 feet**

Ranquear respostas candidatas…

- … caso existam mais de uma.

Q: Who was Queen Victoria's second son?

- Tipo da resposta: Person
- Trecho:

The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

Aprendizagem de máquina

Features para ranquear respostas candidatas

Answer type match: candidata contém uma frase com o tipo de resposta correto

Pattern match: a identidade da expressão regular que casa com a candidata

Question keywords: # de palavras-chave da questão existentes na candidata

Keyword distance: # médio de palavras comuns as palavras-chave da query e da candidata

Novelty factor: pelo menos uma palavra da candidata não está na query

Apposition features: a candidata é um aposto para uma frase contendo termos da questão

Punctuation location: a candidata é imediatamente seguida por uma vírgula, ponto, aspas, ponto e vírgula ou ponto de exclamação.

Sequences of question terms: o comprimento da maior sequência de termos da questão que ocorre na resposta candidata.

IBM Watson: >50 componentes

Métricas de avaliação

1. Accuracy (a resposta casa com a desejada/correta?)
2. *Mean Reciprocal Rank (MRR)*
   - Para cada query, retorna uma lista ranqueada de M respostas candidatas
   - O score da query é 1/Rank da primeira resposta correta
     - se a 1a é correta: 1
     - senão, se a segunda é a correta: ½
     - senão, se a terceira é a correta: ⅓, etc.
     - se nenhuma correta => Score = 0
   - Considere a média sobre todas as N queries

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

Grupo de Pesquisa em
Inteligência e Imagens

Utilizando conhecimento em QA

Extração de relação

- Respostas: bases de dados de relações
  - born-in(“Emma Goldman”, “June 27 1869”)
  - author-of(“Cao Xue Qin”, “Dream of the Red Chamber”)
  - retirado de Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questões: extração de relações em Questões

  Whose granddaughter starred in E.T.?

  (acted-in ?x “E.T.”)

  (granddaughter-of ?x ?y)

Raciocínio temporal

- Bases de dados de relações
  - (e obituários, dicionários biográficos, etc.)
- IBM Watson

  ”In 1594 he took a job as a tax collector in Andalusia”

  Candidatos:
  - Thoreau é uma resposta ruim (nasceu em 1817)
  - Cervantes é possível (estava vivo em 1594)

Conhecimento geospacial

- Beijing é uma boa resposta para ”cidade asiática”
- Califórnia é ”sudoeste de Montana”
- geonames.org:

Grupo de Pesquisa em
Inteligência e Imagens

Sistemas de *Question-Answering (QA)*