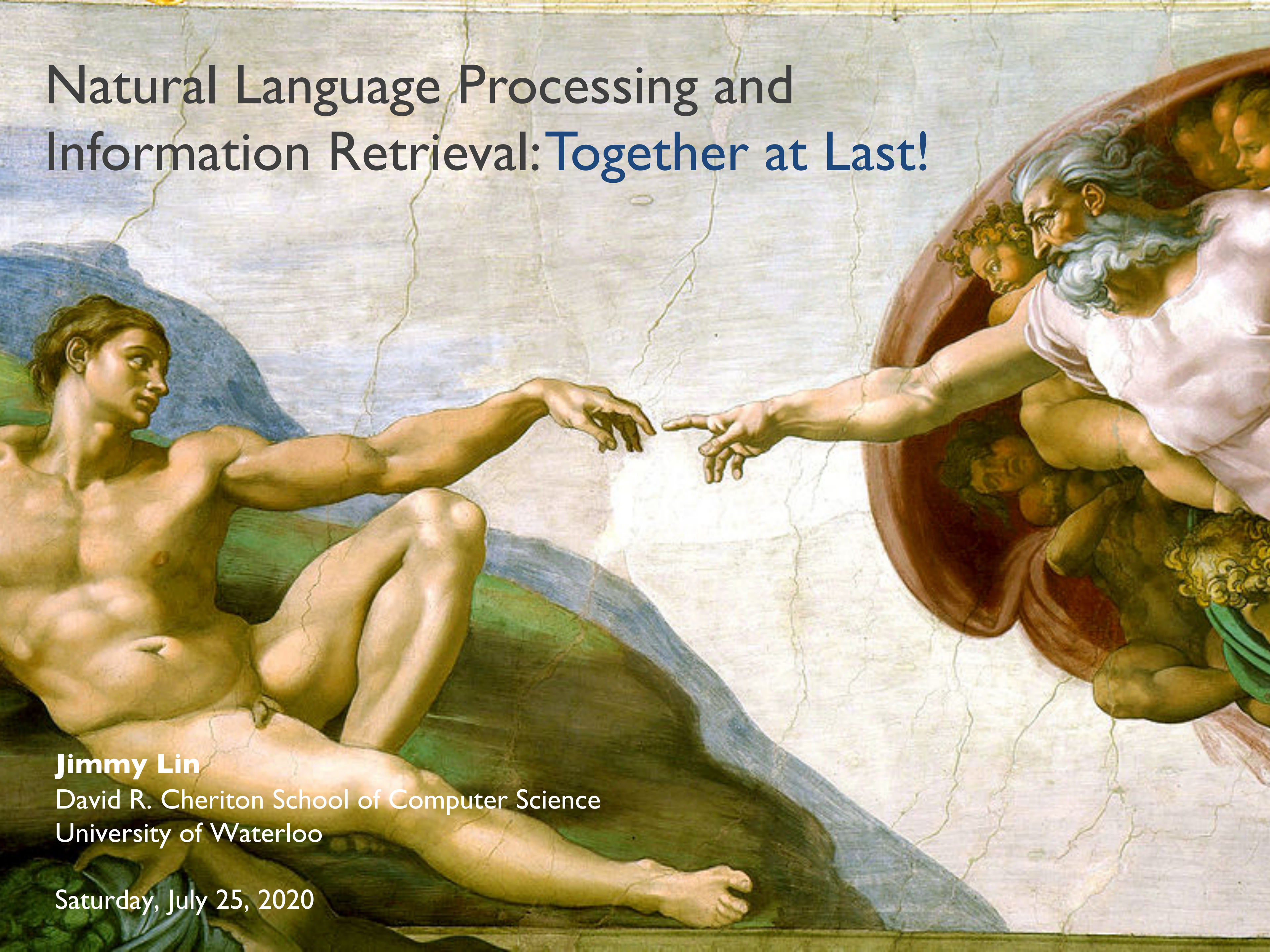
Natural Language Processing and Information Retrieval: Together at Last!

Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo

Saturday, July 25, 2020



It’s an exciting time to do research!

(beginning of a new era…)

It’s an exciting time to do research!

(beginning of a new era…)

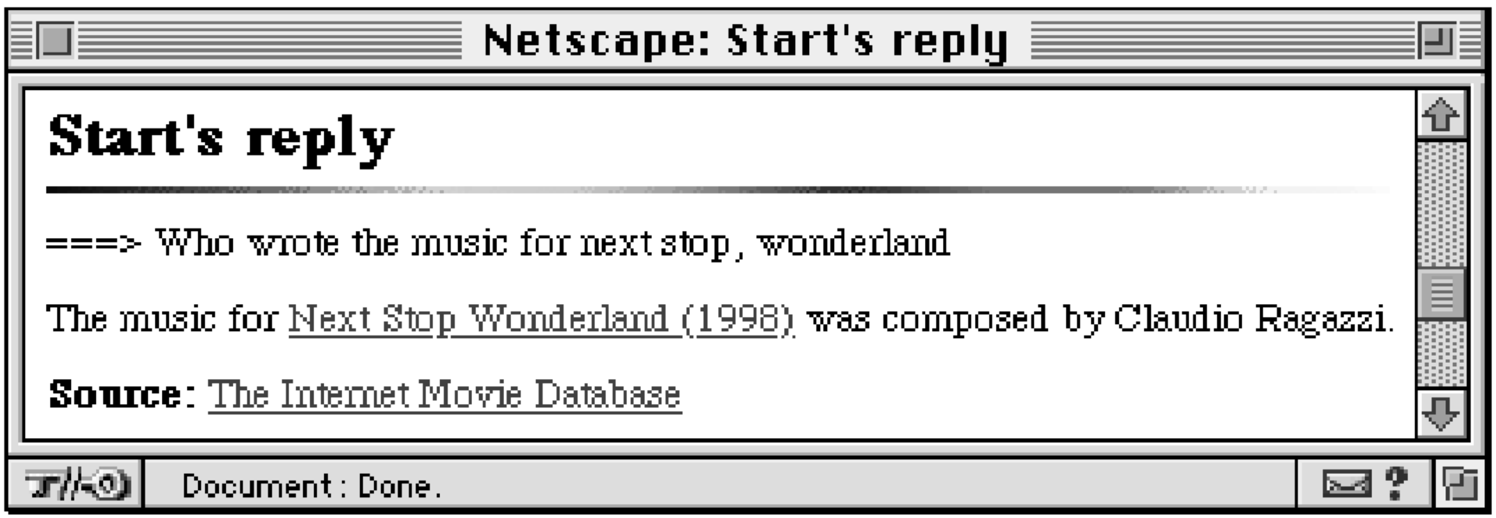
Source: flickr (tapasinthesun/49114923568)

1997: My journey begins

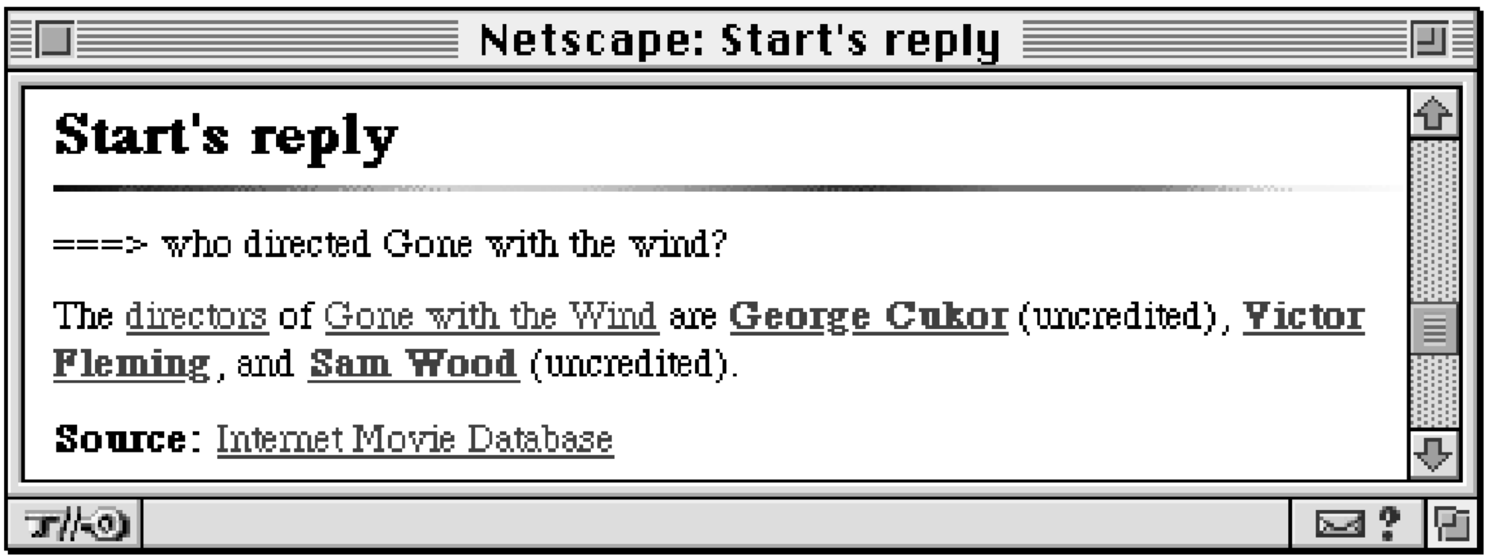
Source: Philip Greenspun, Wikipedia

1993: The START System

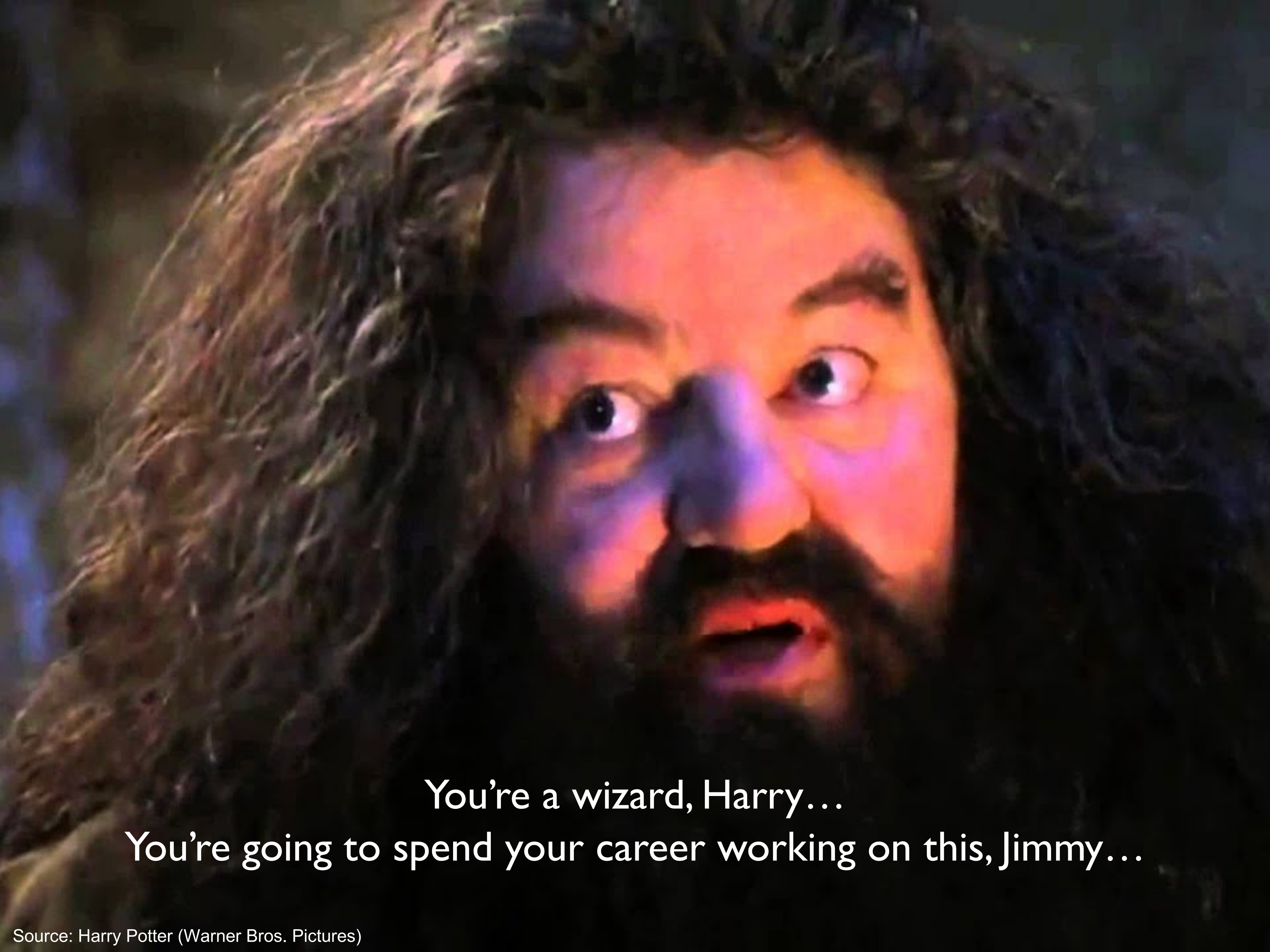
First QA system on the web!



http://start.csail.mit.edu/



http://start.csail.mit.edu/



Source: Harry Potter (Warner Bros. Pictures)

My career-long quest…

Connecting users with relevant information

Source: flickr (71855373@N00/50092059017)

My career-long quest…

Connecting users with relevant information

What? text, speech, images, graphs, semi-structured data, relational data…

Who? general information seekers, domain experts, legal scholars, historians, data scientists, etc.

Information Access

(*ad hoc* retrieval, question answering, summarization, …)

Information Access

The challenge of scale

The challenge of understanding

Working hypothesis:

solving the information access problem requires understanding texts

What does “understanding” mean?

For this talk, I’ll treat it like pornography.

I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description [“hard-core pornography”], and perhaps I could never succeed in intelligibly doing so. But ***I know it when I see it***…

U.S. Supreme Court Justice Potter Stewart
in *Jacobellis v. Ohio* (1964)

counting the frequency of terms
identifying named entities
syntactic parsing
semantic role labeling

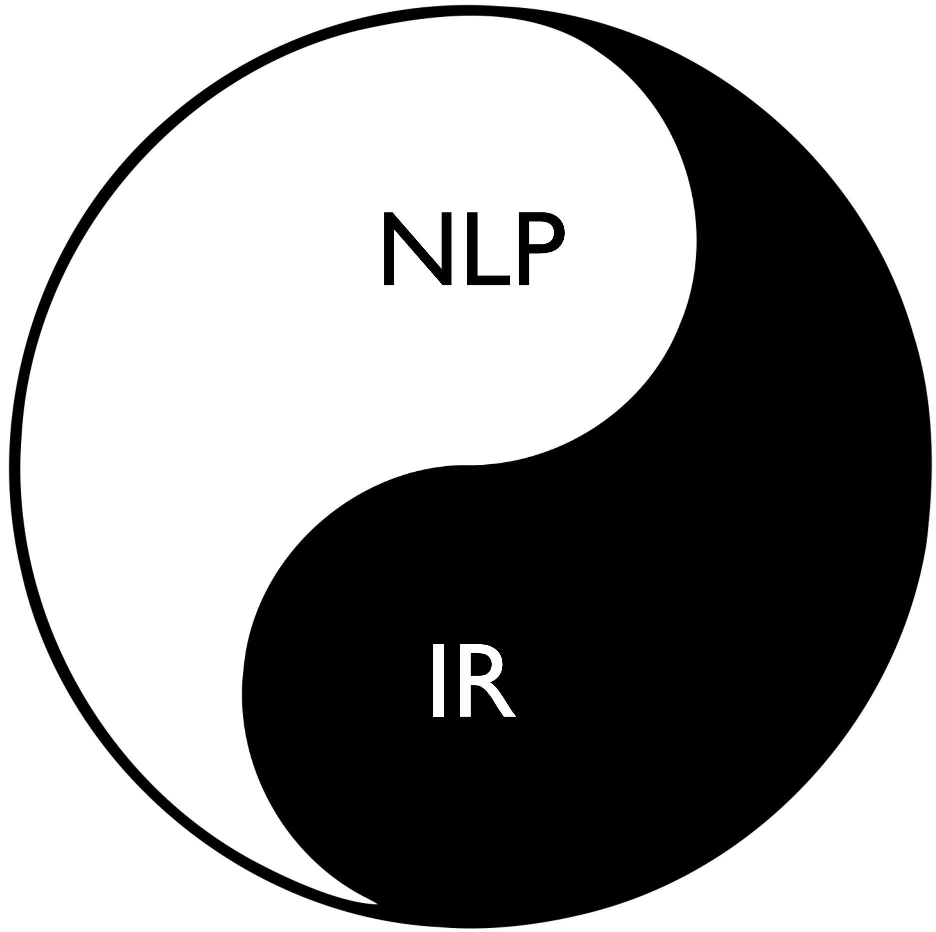
Increasing “understanding”

Working hypothesis:

solving the information access problem
requires understanding texts

Working hypothesis, *revised*:

solving the information access problem
requires the synthesis of NLP and IR



Surely, this must be the case?

Not necessarily so!

Surely, *understanding* bank (side of river) vs. bank (place to deposit money) must help search?

Using WordNet™ to Disambiguate Word Senses for Text Retrieval

Ellen M. Voorhees
Siemens Corporate Research, Inc.
755 College Road East
Princeton, NJ 08540
ellen@learning.scr.siemens.com

Abstract

This paper describes an automatic indexing procedure that uses the "IS-A" relations contained within WordNet and the set of nouns contained in a text to select a sense for each polysemous noun in the text. The result of the indexing procedure is a vector in which some of the terms represent word senses instead of word stems. Retrieval experiments comparing the effectiveness of these sense-based vectors vs. stem-based vectors show the stem-based vectors to be superior overall, although the sense-based vectors do improve the performance of some queries. The overall degradation is due in large part to the difficulty of disambiguating senses in short query statements. An analysis of these results suggests two conclusions: the IS-A links define a generalization/specialization hierarchy that is not sufficient to reliably select the cor-

Nope!

ural language must deal with the problems of polysemy and synonymy. Polysemy, a single word form having more than one meaning, depresses precision by causing false matches, while synonymy, multiple words having the same meaning, depresses recall by causing true conceptual matches to be missed. In principle, polysemy and synonymy can be handled by assigning different senses of a word different *concept identifiers* and assigning the same concept identifier to synonyms. In practice, this requires procedures that are capable of recognizing synonyms, and that can not only detect uses of different senses of a word but can also resolve which meaning is intended in each case.

This paper describes an experiment in which a completely automatic indexing procedure attempts to detect and resolve the senses of the polysemous nouns occurring in the texts of documents and queries. In particular, the procedure selects

SIGIR 1993

Not necessarily so!

Word Sense Disambiguation and Information Retrieval

Mark Sanderson
Department of Computing Science,
University of Glasgow,
Glasgow G12 8QQ
United Kingdom
(email: sanderso@dcs.gla.ac.uk)

Abstract

It has often been thought that word sense ambiguity is a cause of poor performance in Information Retrieval (IR) systems. The belief is that if ambiguous words can be correctly disambiguated, IR performance will increase. However, recent research into the application of a word sense disambiguator to an IR system failed to show any performance increase. From these results it has become clear that more basic research is needed to investigate the relationship between sense ambiguity, disambiguation, and IR.

Using a technique that introduces additional sense ambiguity into a collection, this paper presents research that goes beyond previous work in this field to reveal the influence that ambiguity and disambiguation have on a probabilistic IR system. We conclude that word sense ambiguity is only problematic to an IR system when it is retrieving from very short queries. In addition we argue that if a word sense disambiguator is to be of any use to an IR system, the disambiguator must be able to resolve word senses to a high degree of accuracy.

tl;dr – in principle, would help, but NLP (at the time) sucked too much

1 Introduction

Word ambiguity is not something that we encounter in every day life, except perhaps in the context of jokes. Somehow, when an ambiguous word is spoken in a sentence, we are able to select the correct sense of that word without considering alternative senses. However, in any application where a computer has to process natural language, ambiguity is a problem. For example, if a language translation system encountered the word 'bat' in a

SIGIR 1994

Working hypothesis, *revised*:

solving the information access problem requires the synthesis of NLP and IR

Now let me take you on a journey…


It’s a long and winding road…
(that spans six decades)
Source: flickr (148598741@N02/50125505998)

But you already know where it ends…

Information Access in Two Steps

(1) Select some promising texts

= Tackling the issue of scale

(2) Understand selected texts

= Tackling the issue of understanding

Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

Select some promising texts

Understand selected texts

Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

Select some promising texts

Understand selected texts

(Do we actually need this?)

Working hypothesis, *revised*:
solving the information access problem
requires the synthesis of NLP and IR

Some History

(And yes, NLP and IR existed before neural networks.)

Source: flickr (53292075@N04/48939537413)

Information Access in Two Steps

document (*ad hoc*) retrieval

Select some promising texts

Understand selected texts

(Do we actually need this?)

Information Access in Two Steps

document (*ad hoc*) retrieval

Select some promising texts

U nd
sel exts

Appears not!

Information Access in Two Steps

document (*ad hoc*) retrieval

pre-neural, pre-BERT (pre-history?)

Select some promising texts

Learning to Rank

Li, Hang. Learning to Rank for Information Retrieval and Natural Language Processing. *Morgan & Claypool Publishers,* 2011.

Liu, Tie-Yan. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225-331, 2009.

Information Access in Two Steps

document (*ad hoc*) retrieval

pre-neural, pre-BERT (pre-history?)

Select some promising texts

Learning to Rank

Lots of hand-crafted features, lots of (noisy) data, feed to a supervised model!

(and yes, some of these models were neural networks)

Information Access in Two Steps

document (*ad hoc*) retrieval

Learning to Rank using Gradient Descent

RankNet (ICML, 2005)

Keywords: ranking, gradient descent, neural networks, probabilistic cost functions, internet search

Chris Burges CBURGES@MICROSOFT.COM
Tal Shaked* TAL.SHAKED@GMAIL.COM
Erin Renshaw ERINREN@MICROSOFT.COM
Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399

Ari Lazier ARIEL@MICROSOFT.COM
Matt Deeds MADEEDS@MICROSOFT.COM
Nicole Hamilton NICHAM@MICROSOFT.COM
Greg Hullender GREGHULL@MICROSOFT.COM
Microsoft, One Microsoft Way, Redmond, WA 98052-6399

Lots of hand-crafted features, lots of (noisy) data, feed to a supervised model!

(and yes, some of these models were neural networks)

Abstract

We investigate using gradient descent methods for learning ranking functions; we propose a simple probabilistic cost function, and we introduce RankNet, an implementation of these ideas using a neural network to model the underlying ranking function. We present test results on toy data and on data from a

that maps to the reals (having the model evaluate on pairs would be prohibitively slow for many applications). However (Herbrich et al., 2000) cast the ranking problem as an ordinal regression problem; rank boundaries play a critical role during training, as they do for several other algorithms (Crammer & Singer, 2002; Harrington, 2003). For our application, given that item A appears higher than item B in the out-

Information Access in Two Steps

document (*ad hoc*) retrieval

SIGIR 1993!

Computation of Term Associations by a Neural Network

S.K.M. Wong and Y.J. Cai
Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2

Y.Y. Yao

Department of Mathematical Sciences, Lakehead University
Thunder Bay, Ontario, Canada P7B 5E1

Abstract

This paper suggests a method for computing term associations based on an adaptive bilinear retrieval model. Such a model can be implemented by using a three-layer feedforward neural network. Term associations are modeled by weighted links connecting different neurons, and are derived by the perceptron learning algorithm without the need for introducing any *ad hoc* parameters. The preliminary results indicate the usefulness of neural networks in the design of adaptive information retrieval systems.

The methods for computing term associations can be divided into two categories. One can estimate term relationships directly from the term co-occurrence frequencies. On the other hand, one can infer term associations from the relevance information through feedback. In the first approach, the semantic relationships are derived from the characteristics of term distribution in a document collection (Spark Jones, 1971; van Rijsbergen, 1979; Salton, 1989). These methods are based on the hypothesis that term co-occurrence statistics provide useful information about the relationships between terms. That is, if two or more terms co-occur in many documents, these terms would be more likely semantically related. For example, in the linear associa-

(and yes, some of these models were neural networks)

Information Access in Two Steps

document (*ad hoc*) retrieval

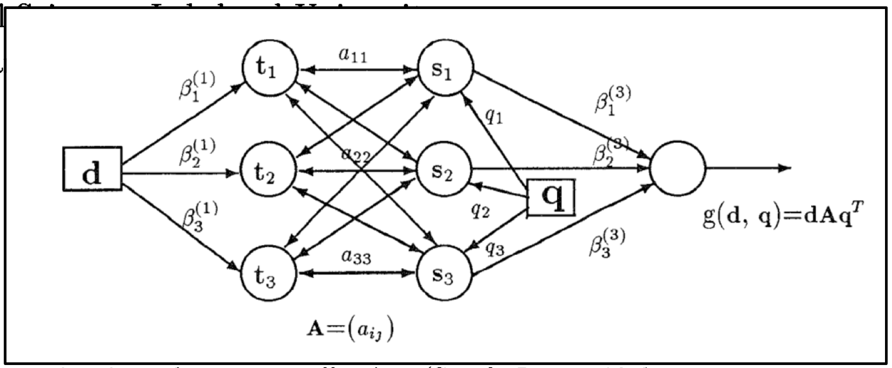
Computation of Term Associations by a Neural Network

S.K.M. Wong and Y.J. Cai
Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2

Y.Y. Yao

Department of Mathematical
Thunder Bay, Onta

SIGIR 1993!



Abstract

This paper suggests a method for computing term associations based on an adaptive bilinear retrieval model. Such a model can be implemented by using a three-layer feed-forward neural network. Term associations are modeled by weighted links connecting different neurons, and are derived by the per-cept... for introducing any *ad hoc* parameters. The preliminary results indicate the usefulness of neural networks in the design of adaptive information retrieval systems.

tion in a document collection (Spark Jones, 1971; van Rijsbergen, 1979; Salton, 1989). These methods are based on the hypothesis that term co-occurrence statistics provide useful information about the relationships between terms. That is, if two or more terms co-occur in many documents, these terms would be more likely semantically related. For example, in the linear associa-

(and yes, some of these models were neural networks)

Information Access in Two Steps

document (*ad hoc*) retrieval

We would call this pointwise learning to rank today!

Optimum Polynomial Retrieval Functions Based on the Probability Ranking Principle

NORBERT FUHR
Technische Hochschule Darmstadt, Darmstadt, West Germany

TOIS 1989!

We show that any approach to developing optimum retrieval functions is based on two kinds of assumptions: first, a certain form of representation for documents and requests, and second, additional simplifying assumptions that predefine the type of the retrieval function. Then we describe an approach for the development of optimum polynomial retrieval functions: request-document pairs (f_l, d_m) are mapped onto description vectors $\vec{x}(f_l, d_m)$, and a polynomial function $e(\vec{x})$ is developed such that it yields estimates of the probability of relevance $P(R \mid \vec{x}(f_l, d_m))$ with minimum square errors. We give experimental results for the application of this approach to documents with weighted indexing as well as to documents with complex representations. In contrast to other probabilistic models, our approach yields estimates of the actual probabilities, it can handle very complex representations of documents and requests, and it can be easily applied to multivalued relevance scales. On the other hand, this approach is not suited to log-linear probabilistic models and it needs large samples of relevance feedback data for its application.

Categories and Subject Descriptors: G.1.2 [**Numerical Analysis**]: Approximation—*least squares approximation*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models*

General Terms: Experimentation, Theory

Information Access in Two Steps

document (*ad hoc*) retrieval

We would call this pointwise learning to rank today!

Optimum Polynomial Retrieval Functions
Based on the Probability Ranking Principle

NORBERT FUHR
Technische Hochschule Darmstadt, Darmstadt, West Germany

TOIS 1989!

We show that any approach to developing optimu
assumptions: first, a certain form of representation fo
simplifying assumptions that predefine the type o
approach for the development of optimum polynor
(f_l, d_m) are mapped onto description vectors $\vec{x}(f_l, d$
such that it yields estimates of the probability of
errors. We give experimental results for the applicat
indexing as well as to documents with complex rep
models, our approach yields estimates of the act
representations of documents and requests, and it
scales. On the other hand, this approach is not suite
large samples of relevance feedback data for its appl

Categories and Subject Descriptors: G.1.2 [**Nume**
approximation; H.3.1 [**Information Storage and**
indexing methods; H.3.3 [**Information Storag**
Retrieval—*retrieval models*

Table VI. Elements of the Description Vector $\vec{x}(f_l, d_m)$

Element	Description
x_1	number of descriptors common to query and document
x_2	log(number of descriptors common to query and document)
x_3	highest indexing weight of a common descriptor
x_4	lowest indexing weight of a common descriptor
x_5	number of common descriptors with weight ≥ 0.15
x_6	number of noncommon descriptors with weight ≥ 0.15
x_7	number of descriptors in the document with weight ≥ 0.15
x_8	$\log \sum$ (indexing weights of common descriptors)
x_9	log(number of descriptors in the query)
x_{10}	log(min(size of output set, 100))
x_{11}	$= 1$, if size of output set > 100
x_{12}	$= 1$, if request is about nuclear physics
x_{13}	proportion of relevant documents in the output set

Information Access in Two Steps

document (*ad hoc*) retrieval

Select some promising texts

Understanding, Smunderstanding!

Learning to Rank

Working hypothesis, *revised*:

solving the information access problem requires the synthesis of NLP and IR

For *ad hoc* retrieval, particularly at scale?

Reject!

Information Access in Two Steps

question answering

Need fine-grained analysis – the perfect setup!

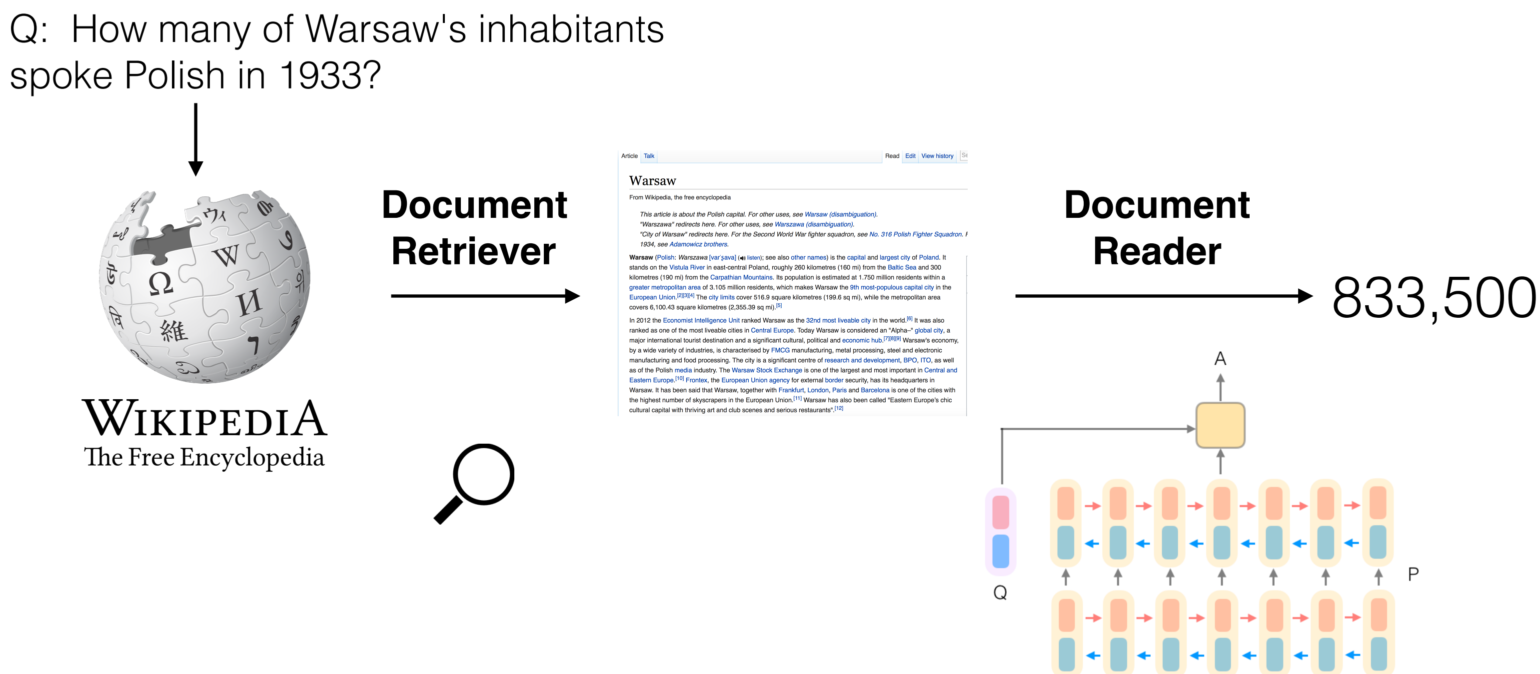
Select some promising texts

Understand selected texts

(Do we actually need this?)

Reading Wikipedia to Answer Open-Domain Questions

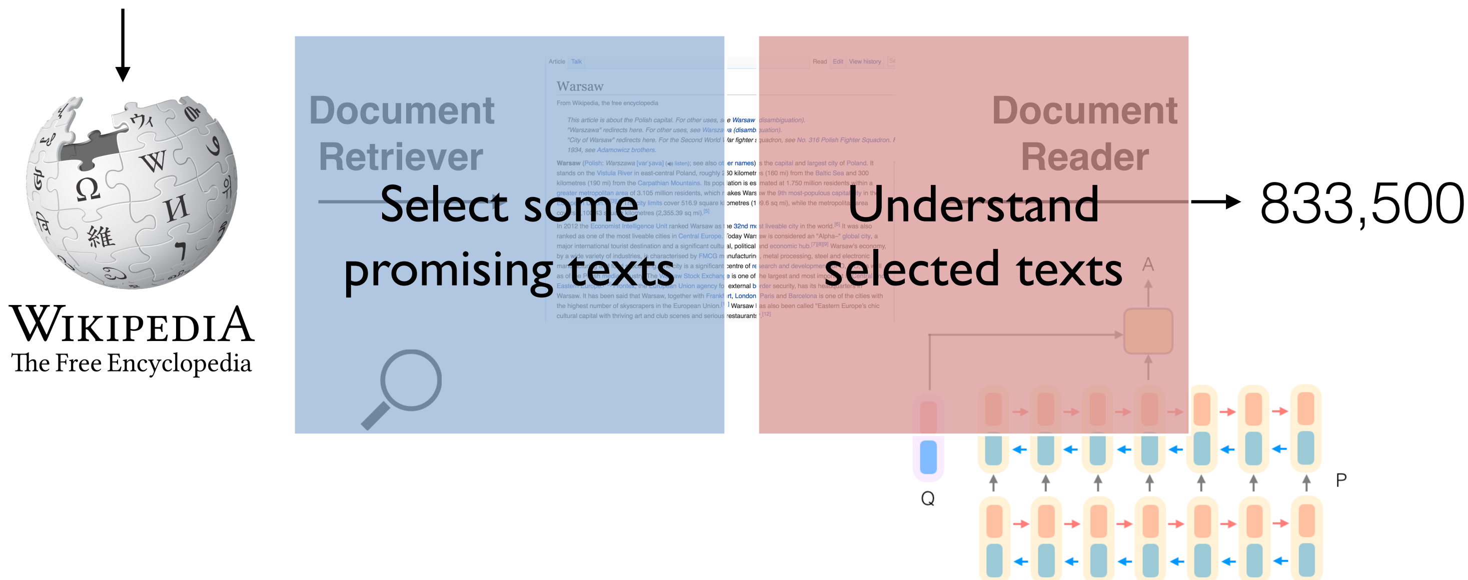
Chen et al. (ACL 2017)

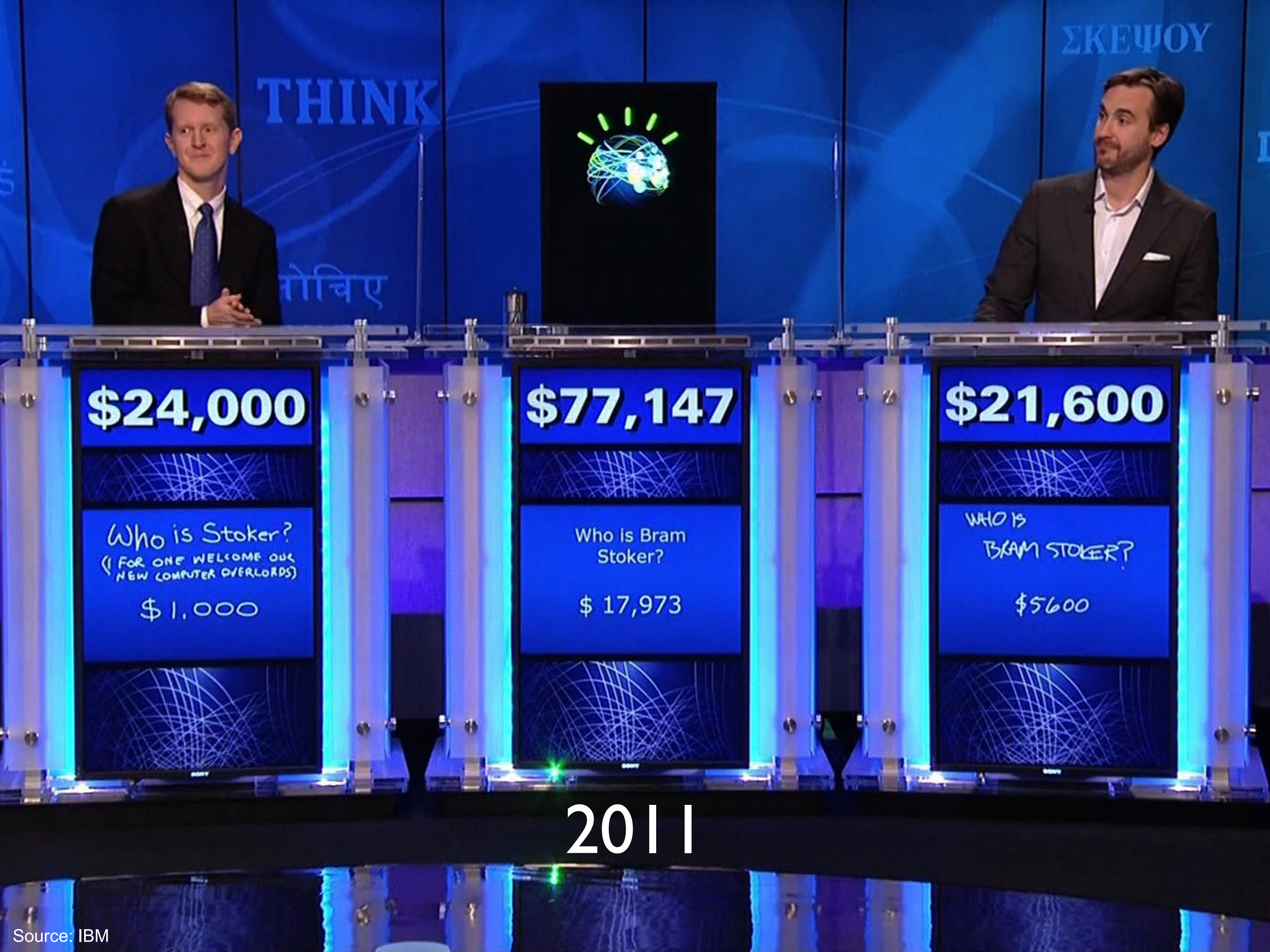


Reading Wikipedia to Answer Open-Domain Questions

Chen et al. (ACL 2017)

Q: How many of Warsaw's inhabitants spoke Polish in 1933?





2011

Source: IBM

TREC-8 (1999)

The TREC-8 Question Answering Track Evaluation

Ellen M. Voorhees, Dawn M. Tice
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The TREC-8 Question Answering track was the first large-scale evaluation of systems that return answers, as opposed to lists of documents, in response to a question. As a first evaluation, it is important to examine the evaluation methodology itself to understand any limits on the conclusions that can be drawn from the evaluation and possibly to find ways to improve subsequent evaluations. This paper has two main goals: to describe in detail how the evaluation was implemented, and to examine the consequences of the methodology on the comparative performance of the systems participating in the evaluation. The examination uncovered no serious flaws in the methodology, supporting its continued use for question answering evaluation. Nonetheless, redefining the specific task to be performed so that it more closely matches an actual user task does appear warranted.

1 Introduction

The Text REtrieval Conference (TREC) is a series of workshops designed to advance the state-of-the-art in text retrieval by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. Evaluating competing technologies on a common test set has had the desired effect of increasing text retrieval system effectiveness as demonstrated, for example, by the doubling of performance of the SMART system since the beginning of TREC [1]. However, users generally would prefer to receive *answers* in response to their questions, as opposed to the document lists traditionally returned by text retrieval systems. The TREC-8 Question Answering Track is an initial effort to bring the benefits of large-scale evaluation to bear on the question answering task.

QA in the early 2000s

The Use of External Knowledge in Factoid QA

Eduard Hovy, Ulf Hermjakob, Chin-Yew Lin

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
tel: 310-448-8731
fax: 310-823-6714
email: {hovy,ulf,cyl}@isi.edu

Abstract

This paper describes recent development in the Webclopedia QA system, focusing on the use of knowledge resources such as WordNet and a QA typology to improve the basic operations of candidate answer retrieval, ranking, and answer matching.

1. Introduction

The Webclopedia factoid QA system increasingly makes use of syntactic and semantic (world) knowledge to improve the accuracy of its results. Previous TREC QA evaluations made clear the need for using such external knowledge to improve answers. For example, for definition-type questions such as

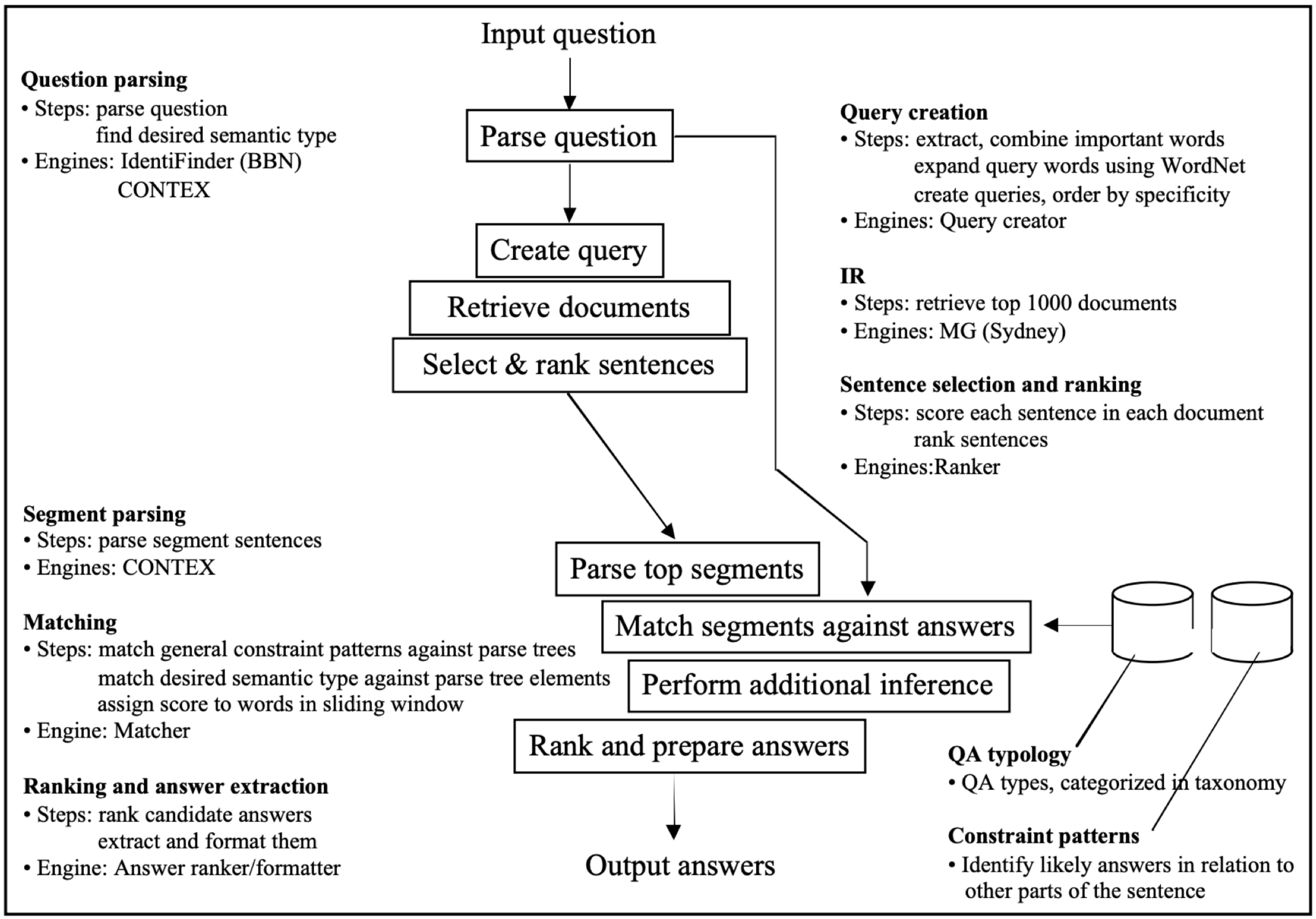
Q: what is bandwidth?

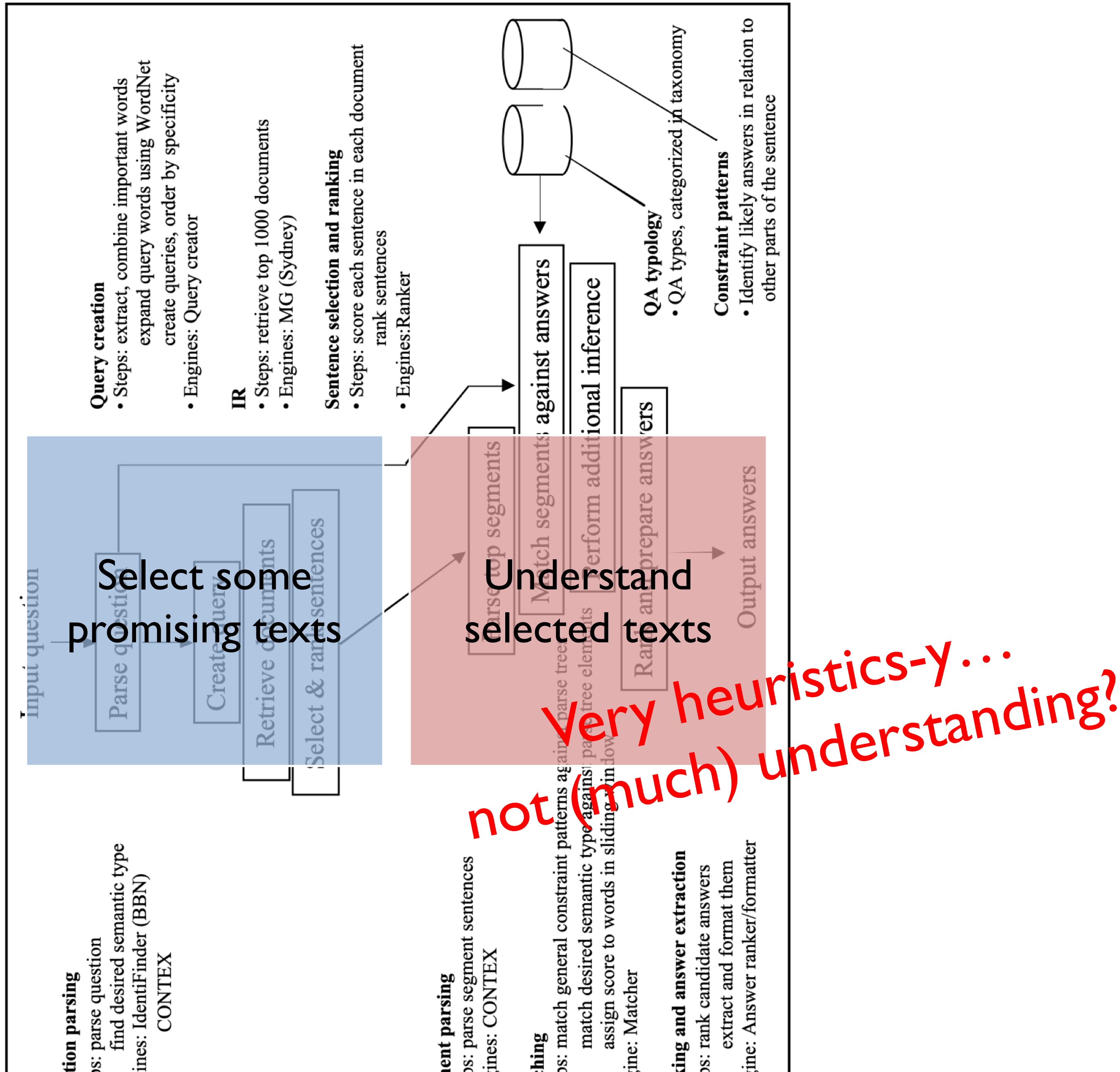
the system uses WordNet to extract words used in the term definitions before searching for definitions in the answer corpus, and boosts candidate answer scores appropriately. Such definitional WordNet glosses have helped definition answers (10% for definition questions, which translates to about 2% overall score in the TREC-10 QA evaluation, given that as many as a little over 100 out of 500 TREC-10 questions were definition questions).

This knowledge is of one of two principal types: generic knowledge about language, and knowledge about the world. After outlining the general system architecture, this paper describes the use of knowledge to

TREC 2001

QA in the early 2000s




Select some promising texts
Understand selected texts
Very heuristics-y…
not (much) understanding?
Query creation
• Steps: extract, combine important words
expand query words using WordNet
create queries, order by specificity
• Engines: Query creator
IR
• Steps: retrieve top 1000 documents
• Engines: MG (Sydney)
Sentence selection and ranking
• Steps: score each sentence in each document
rank sentences
• Engines:Ranker
QA typology
• QA types, categorized in taxonomy
Constraint patterns
• Identify likely answers in relation to
other parts of the sentence
Match segments against answers
Perform additional inference
Output answers

QA in the early 2000s

Data-Intensive Question Answering

Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais and Andrew Ng
Microsoft Research
One Microsoft Way
Redmond, WA 98052
{brill, mbanko, sdumais}@microsoft.com
jlin@ai.mit.edu; ang@cs.berkeley.edu

1 Introduction

Microsoft Research Redmond participated for the first time in TREC this year, focusing on the question answering track. There is a separate report in this volume on the Microsoft Research Cambridge submissions for the filtering and Web tracks (Robertson et al., 2002). We have been exploring data-driven techniques for Web question answering, and modified our system somewhat for participation in TREC QA. We submitted two runs for the main QA track (AskMSR and AskMSR2).

Data-driven methods have proven to be powerful techniques for natural language processing. It is still unclear to what extent this success can be attributed to specific techniques, versus simply the data itself. For example, Banko and Brill (2001) demonstrated that for confusion set disambiguation, a prototypical disambiguation-in-string-context problem, the amount of data used far dominates the learning method employed in improving labeling accuracy. The more training data that is used, the greater the chance that a new sample being processed can be trivially related

TREC 2001

AskMSR

Select some promising texts

Understanding, Smunderstanding!

Understand selected texts

AskMSR

Select some promising texts

Count *n*-grams

Certainly no understanding!

Bill Gates to Keynote International Joint Conference on Artificial Intelligence

August 6, 2001 |

SEATTLE, Aug. 6, 2001 — Microsoft Corp. Chairman and Chief Software Architect Bill Gates is scheduled to deliver the keynote presentation at the International Joint Conference on Artificial Intelligence (IJCAI) tomorrow morning at the Washington State Convention Center in Seattle. IJCAI is the main international conference on artificial intelligence, held biennially, but only once every four years in North America. Gates' speech,
"AI in the Computing Experience: Challenges and Opportunities,"
will address key challenges and opportunities for enhancing the computer user experience with innovations that leverage developments in artificial intelligence.

…

- **AskMSR.** Automated question answering from information on the World Wide Web (Eric Brill, Machine Learning and Applied Statistics Group, Microsoft Research)

Bill Gates to Keynote International Joint Conference on Artificial Intelligence

August 6, 2001 |

SEATTLE, Aug. 6, 2001 — Microsoft Corp. Chairman and C
the keynote presentation at the International Joint Confer
the Washington State Convention Center in Seattle. IJCAI
intelligence, held biennially, but only once every four yea
"AI in the Computing Experience: Challenges and Opport
will address key challenges and opportunities for enhanc
leverage developments in artificial intelligence.

IJCAI Keynote Speaker Bill Gates. Photograph by Andrew Buchanan.

…

- **AskMSR.** Automated question answering from information on the World Wide Web Eric Brill, Machine Learning and Applied Statistics Group, Microsoft Research)

AskMSR

Select some promising texts

Count
n-grams

… not very satisfying

My master's thesis

Selectively Using Relations to Improve Precision in Question Answering

Boris Katz and **Jimmy Lin**
MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139
{boris,jimmylin}@ai.mit.edu

Abstract

Despite the intuition that linguistically sophisticated techniques should be beneficial to question answering, real gains in performance have yet to be demonstrated empirically in a reliable manner. Systems built around sophisticated linguistic analysis generally perform worse than their linguistically-uninformed cousins. We believe that the key to effective application of natural language processing technology is to selectively employ it

cess, there exist empirical limits on the effectiveness of this approach. By analyzing a subset of TREC-9 and CBC questions, Light et al. (2001) established an expected upper bound on the performance of a question answering system with perfect passage retrieval, named-entity detection, and question classification at around 70%. The primary reason for this limit is that many named entities of the same semantic type often occur close together, and a QA system, without the aid

EACL Workshop 2003

My master's thesis

(Q1) What do frogs eat?

(A1) Adult *frogs eat* mainly insects and other small animals, including earthworms, minnows, and spiders.

(A2) Alligators *eat* many kinds of small animals that live in or near the water, including fish, snakes, *frogs*, turtles, small mammals, and birds.

(A3) Some bats catch fish with their claws, and a few species *eat* lizards, rodents, small birds, tree *frogs*, and other bats.

(Q2) What is the largest volcano in the Solar System?

(B1) Mars boasts many extreme geographic features; for example, Olympus Mons, the *largest volcano in the solar system*.

(B2) The Galileo probe's mission to Jupiter, the *largest* planet *in the Solar system*, included amazing photographs of the *volcanoes* on Io, one of its four most famous moons.

(B3) Even the *largest volcanoes* found on Earth are puny in comparison to others found around our own cosmic backyard, *the Solar System*.

(B4) Olympus Mons, which spans an area the size of Arizona, is the *largest volcano in the Solar System*.

ʀelations to Improve Precision
estion Answering

Katz
ial Int
Techn
ıbridg
immy

(1) [bird eat snake]
(1′) [snake eat bird]
(2) [largest `adjmod` planet]
[planet `poss` volcano]
(2′) [largest `adjmod` volcano]
[planet `poss` volcano]
(3) [house by river]
(3′) [river by house]
(4) [Germans defeat French]
(4′) [French defeat Germans]

phisti-
ıestion
yet to
man-
nguis-
ı their
ve that
the key to effective application of natural language
processing technology is to selectively employ it

formance of a question answering system with perfect passage retrieval, named-entity detection, and question classification at around 70%. The primary reason for this limit is that many named entities of the same semantic type often occur close together, and a QA system, without the aid of any additional knowledge, would be forced to

EACL Workshop 2003

My master's thesis

Select some promising texts

Match linguistic relations

Closer to understanding?

START

Annotating the World Wide Web using Natural Language

Boris Katz
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
545 Technology Square
Cambridge, MA 02139, USA
boris@ai.mit.edu

This paper describes the START Information Server built at the MIT Artificial Intelligence Laboratory. Available on the World Wide Web since December 1993, the START Server provides users with access to multi-media information in response to questions formulated in English. Over the last 3 years, the START Server answered hundreds of thousands of questions from users all over the world.

The START Server is built on two foundations: the sentence-level Natural Language processing capability provided by the START Natural Language system (Katz [1990]) and the idea of natural language annotations for multi-media information segments. This paper starts with an overview of sentence-level processing in the START system and then explains how annotating information segments with collections of English sentences makes it possible to use the power of sentence-level natural language processing in the service of multi-media information access. The paper ends with a proposal to annotate the World Wide Web.

An Overview of the START system

The START natural language system (SynTactic Anal-

Given an English sentence containing various relative clauses, appositions, multiple levels of embedding, *etc*, the START system first breaks it up into smaller units, called *kernel* sentences (usually containing one verb). After separately analyzing each kernel sentence, START rearranges the elements of all parse trees it constructs into a set of embedded representational structures. These structures are made up of a number of fields corresponding to various syntactic parameters of a sentence, but the three most salient parameters, the subject of a sentence, the object, and the relation between them are singled out as playing a special role in indexing. These parameters are explicitly represented in a discrimination network for efficient retrieval. As a result, all sentences analyzed by START are indexed as embedded *ternary expressions* *(T-expressions)*, <**subject relation object**>. Certain other parameters (adjectives, possessive nouns, prepositional phrases, *etc.*) are used to create additional T-expressions in which prepositions and several special words may serve as relations. For instance, the following simple sentence

RIAO 1997

START

Select some promising texts

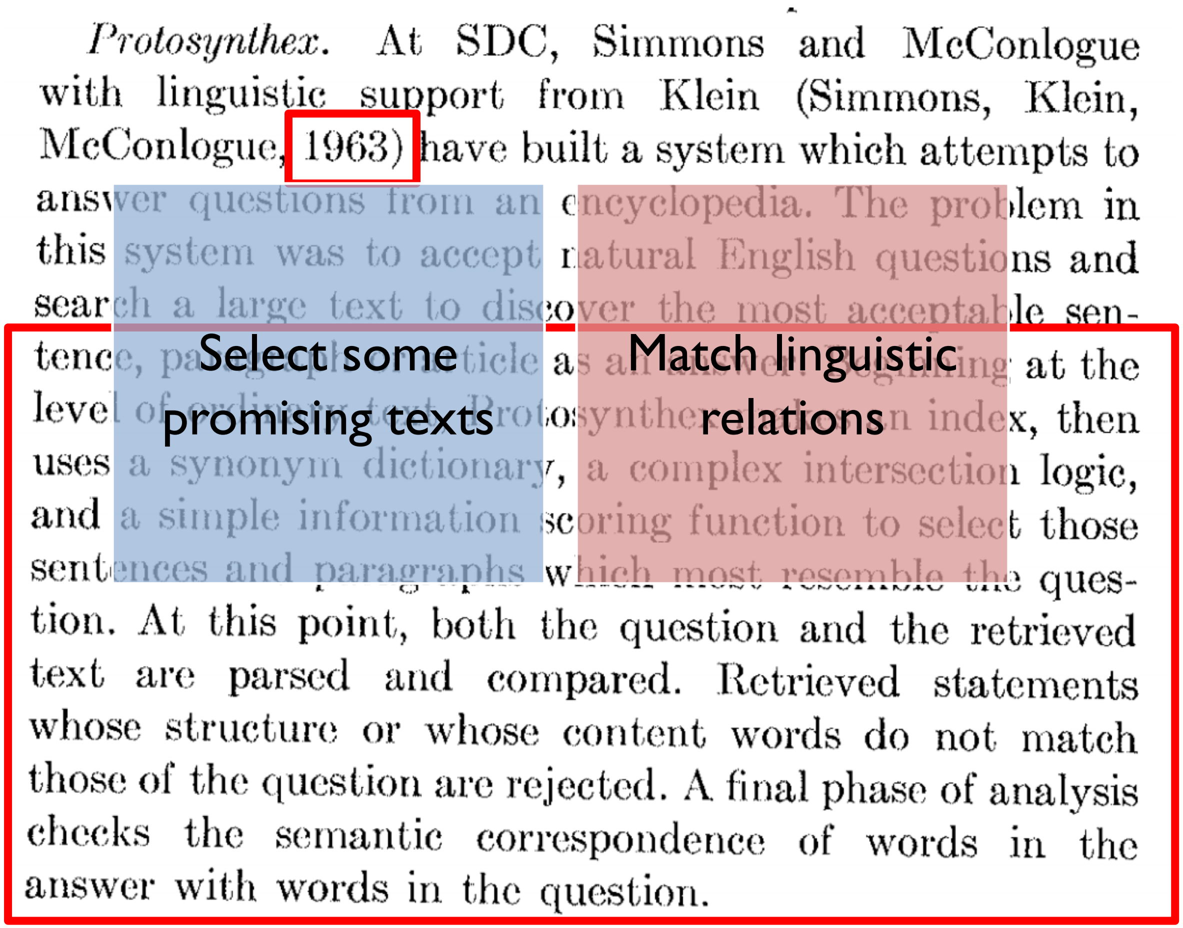
Match linguistic relations

Protosynthex

Protosynthex. At SDC, Simmons and McConlogue with linguistic support from Klein (Simmons, Klein, McConlogue, 1963) have built a system which attempts to answer questions from an encyclopedia. The problem in this system was to accept natural English questions and search a large text to discover the most acceptable sentence, paragraph or article as an answer. Beginning at the level of ordinary text, Protosynthex makes an index, then uses a synonym dictionary, a complex intersection logic, and a simple information scoring function to select those sentences and paragraphs which most resemble the question. At this point, both the question and the retrieved text are parsed and compared. Retrieved statements whose structure or whose content words do not match those of the question are rejected. A final phase of analysis checks the semantic correspondence of words in the answer with words in the question.

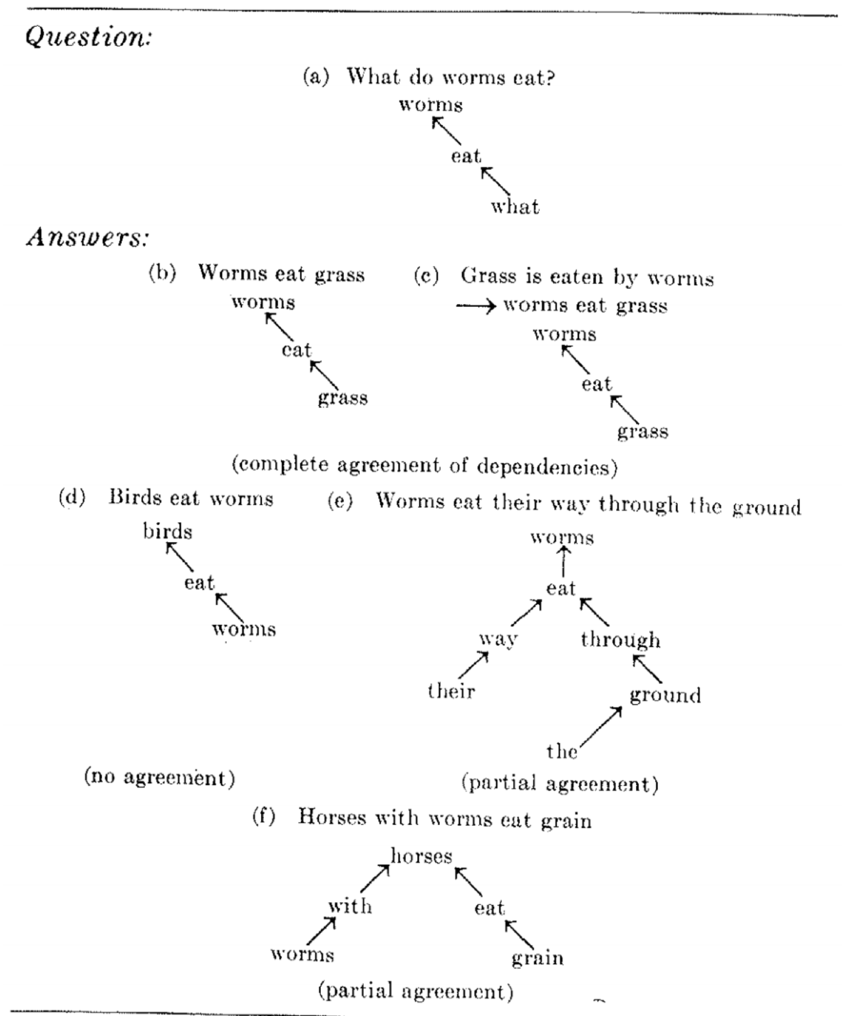
Simmons. Answering English Questions by Computer: A Survey. *CACM*, 8(1):53-70, 1965.

Protosynthex


Protosynthex. At SDC, Simmons and McConlogue with linguistic support from Klein (Simmons, Klein, McConlogue, 1963) have built a system which attempts to answer questions from an encyclopedia. The problem in this system was to accept natural English questions and search a large text to discover the most acceptable sentence, paragraph, or article as an answer. Beginning at the level of ordinary text, Protosynthex makes an index, then uses a synonym dictionary, a complex intersection logic, and a simple information scoring function to select those sentences and paragraphs which most resemble the question. At this point, both the question and the retrieved text are parsed and compared. Retrieved statements whose structure or whose content words do not match those of the question are rejected. A final phase of analysis checks the semantic correspondence of words in the answer with words in the question.

Simmons. Answering English Questions by Computer: A Survey. *CACM*, 8(1):53-70, 1965.

Protosynthex



Simmons. Answering English Questions by Computer: A Survey. *CACM*, 8(1):53-70, 1965.

Information Access in Two Steps

question answering

Select some promising texts

Match linguistic relations

Affirmed?

Working hypothesis, *revised*:

solving the information access problem
requires the synthesis of NLP and IR

Information Access in Two Steps

question answering

Select some promising texts

Match linguistic relations

Unfortunately, none of this really worked… robustly ☹

Until it *finally* worked…

Rank Learning for Factoid Question Answering with Linguistic and Semantic Constraints

Matthew W. Bilotti, Jonathan Elsas, Jaime Carbonell and Eric Nyberg
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213, USA
{ mbilotti, jelsas, jgc, ehn }@cs.cmu.edu

ABSTRACT

This work presents a general rank-learning framework for passage ranking within Question Answering (QA) systems using linguistic and semantic features. The framework enables query-time checking of complex linguistic and semantic constraints over keywords. Constraints are composed of a mixture of keyword and named entity features, as well as features derived from semantic role labeling. The framework supports the checking of constraints of arbitrary length relating any number of keywords. We show that a trained ranking model using this rich feature set achieves greater than a 20% improvement in Mean Average Precision over baseline keyword retrieval models. We also show that constraints based on semantic role labeling features are particularly effective for passage retrieval; when they can be leveraged, an 40% improvement in MAP over the baseline can be realized.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

question, on the front end, and post-retrieval, to locate answers among the results. If QA systems are ever to become competitive with the *ad hoc* keyword search engines that are ubiquitous in the lives of today's internet users, both latency and accuracy must be improved. Both of these goals can be addressed by improving the quality of the embedded passage retrieval component.

Poor passage retrieval quality within QA systems stems in part from a mismatch between what the system wants and what the embedded retrieval component is able to query. Internally, QA systems represent their *information needs* as sets of linguistic and semantic constraints that a retrieved passage must satisfy if it answers the question. Many passage retrieval approaches commonly used in QA systems can not check these types of constraints at query time. As a result, QA systems are forced to approximate their information needs in terms of classic *ad hoc* retrieval primitives such as bag-of-words, proximity and named entity features.

For many questions, the classic feature set poorly approximates the information need, resulting in the retrieval of too few answer-bearing passages and/or too many false positives. This degradation in passage retrieval quality overburdens the downstream Answer Generation component, which must determine whether each retrieved passage is answer-

CIKM 2010

Until it *finally* worked…

Rank Learning for Factoid Question Answering with Linguistic and Semantic Constraints

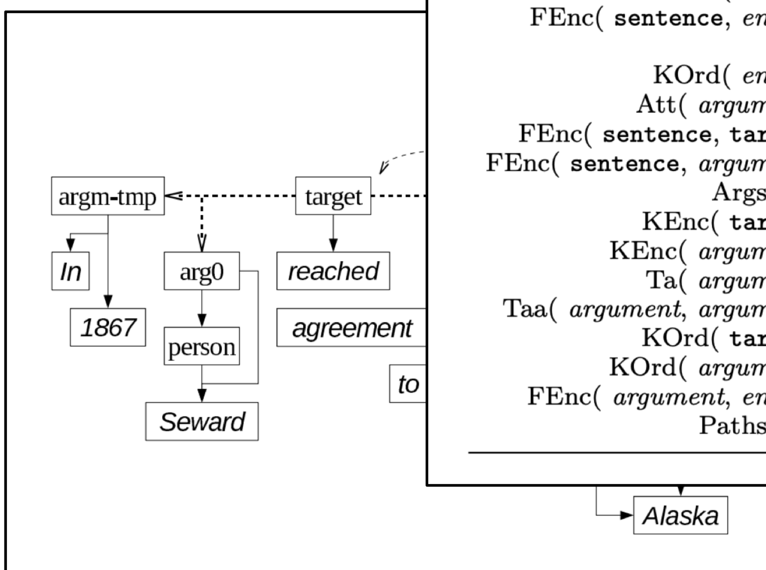
Matth… …ell and Eric Nyberg

…u.edu

… featurized

Feature Name	Groups
Baseline retrieval score	1-8
KEnc(**sentence**)	1,2,6,7,8
KOrd(**sentence**)	2,4,7,8
KEnc(*entity*)	3,4,8
FEnc(**sentence**, *entity*)	3,4,8
Ans	3,4,8
KOrd(*entity*)	4,8
Att(*argument*)	5,6,7,8
FEnc(**sentence**, **target**)	5,6,7,8
FEnc(**sentence**, *argument*)	5,6,7,8
Args(N)	5,6,7,8
KEnc(**target**)	6,7,8
KEnc(*argument*)	6,7,8
Ta(*argument*)	6,7,8
Taa(*argument*, *argument*)	6,7,8
KOrd(**target**)	7,8
KOrd(*argument*)	7,8
FEnc(*argument*, *entity*)	8
Paths(N)	8

Linguistic relations



Input: Number of passage pairs to sample T, Committee size N_{com}, List of training relevant/non-relevant passage pairs $S = R \times N = \{(\mathbf{p}_{nq}, \mathbf{p}_{rq})\}$ **Output:** Set of feature weight vectors and their success counters $K = \{(\mathbf{w}^k, c_k) | k = 1 \dots N_{\text{com}}\}$

1. Initialize $i = 0$, success counter $c_i = 0$, initial parameters $\mathbf{w}^0$, committee $K = \emptyset$.
2. For $t = 0, \dots, T$:

   From S, sample query q and relevant/non-relevant passages $(\mathbf{p}_{nq}, \mathbf{p}_{rq})$

   If $Score(\mathbf{p}_{nq}, \mathbf{w}^i) \geq Score(\mathbf{p}_{rq}, \mathbf{w}^i)$ then

   $(\mathbf{w}_{\min}, c_{\min}) \in K$ s.t. $c_{\min} = \min_k c_k \in K$

   If $c_i > c_{\min}$ then: add $(\mathbf{w}^i, c_i)$ to K

   If $|K| > N_{sub}$: remove $(\mathbf{w}_{\min}, c_{\min})$ from K

   update: $\mathbf{w}^{i+1} = \mathbf{w}^i + (\mathbf{p}_{rq} - \mathbf{p}_{nq})$ and $i = i+1$

   Else update: $c_i = c_i + 1$
3. Output: K

… fed to a ML ranker

tion ne…
as bag…
For…
imates the information need, resulting in the retrieval of too few answer-bearing passages and/or too many false positives. This degradation in passage retrieval quality overburdens the downstream Answer Generation component, which must determine whether each retrieved passage is answer-

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

CIKM 2010

Until it *finally* worked…

Select some promising texts

Learning to Rank

So it's basically this…

with hand-crafted linguistic features

Working hypothesis, *revised*:

solving the information access problem requires the synthesis of NLP and IR

For question answering?

Little novelty.
Reject!

The ending you know is coming…

The beginning of the BERT craze!
January 2019

MS MARCO Leaderboard

MS MARCO

Passage Retrieval Task

Model	Ranking Style	Submission Date	MRR@10 On Eval	MRR@10 On Dev
BERT + Small Training Rodrigo Nogueira(1) and Kyunghyun Cho(2) - New York University(1,2), Facebook AI Research(2) [Nogueira, et al. '19] and [Code]	ReRanking	January 7th, 2019	0.359 +30%	0.365
IRNet (Deep CNN/IR Hybrid Network) Dave DeBarr, Navendu Jain, Robert Sim, Justin Wang, Nirupama Chandrasekaran – Microsoft	ReRanking	January 2nd, 2019	0.281	0.278

PASSAGE RE-RANKING WITH BERT

arXiv:1901.04085, 2019.

Rodrigo Nogueira
New York University
rodrigonogueira@nyu.edu

Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

ABSTRACT

Recently, neural models pretrained on a language modeling task, such as ELMo (Peters et al., 2017), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2018), have achieved impressive results on various natural language processing tasks such as question-answering and natural language inference. In this paper, we describe a simple re-implementation of BERT for query-based passage re-ranking. Our system is the state of the art on the TREC-CAR dataset and the top entry in the leaderboard of the MS MARCO passage retrieval task, outperforming the previous state of the art by 27% (relative) in MRR@10. The code to reproduce our results is available at https://github.com/nyu-dl/dl4marco-bert

https://microsoft.github.io/msmarco/

Information Access in Two Steps

(almost) document (*ad hoc*) retrieval

Select some promising texts

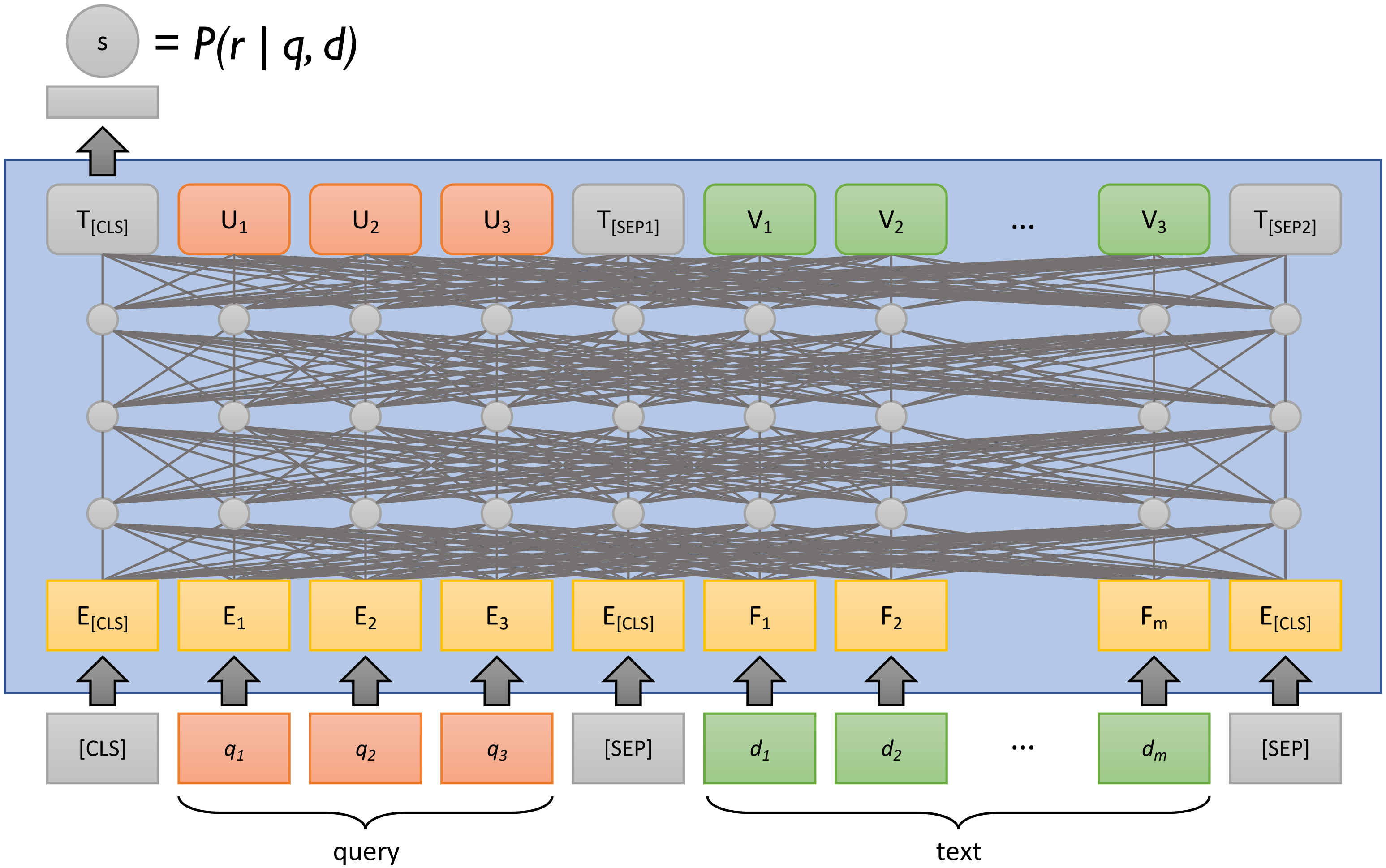
Understand selected texts

Information Access in Two Steps

(almost) document (*ad hoc*) retrieval

Select some promising texts

+



s
= P(r | q, d)
T[CLS]
U1
U2
U3
T[SEP1]
V1
V2
...
V3
T[SEP2]
E[CLS]
E1
E2
E3
E[CLS]
F1
F2
Fm
E[CLS]
[CLS]
q1
q2
q3
[SEP]
d1
d2
...
dm
[SEP]
query
text

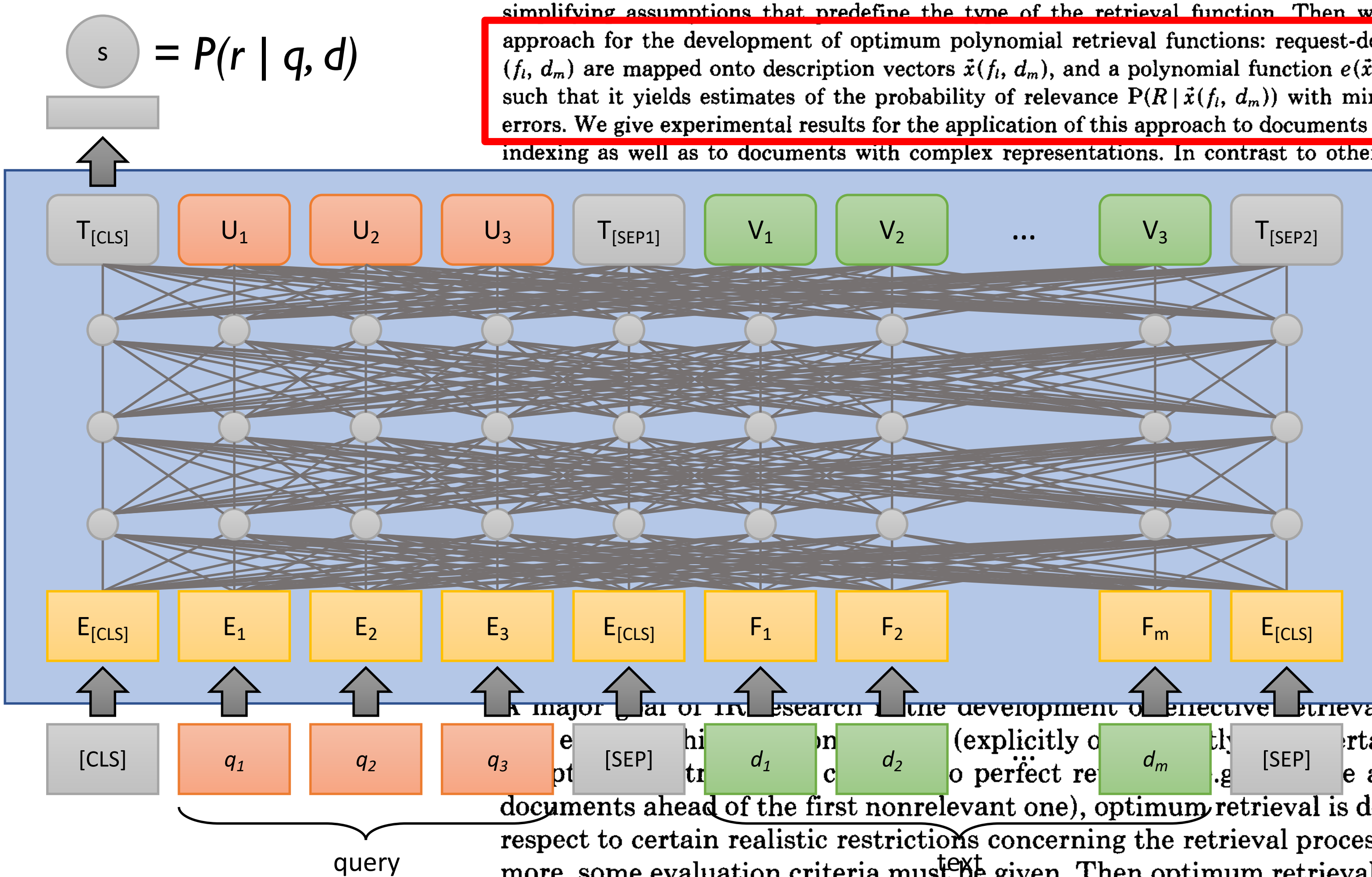
Optimum Polynomial Retrieval Functions Based on the Probability Ranking Principle

NORBERT FUHR
Technische Hochschule Darmstadt, Darmstadt, West Germany

We show that any approach to developing optimum retrieval functions is based on two kinds of assumptions: first, a certain form of representation for documents and requests, and second, additional simplifying assumptions that predefine the type of the retrieval function. Then we describe an approach for the development of optimum polynomial retrieval functions: request-document pairs (f_l, d_m) are mapped onto description vectors $\vec{x}(f_l, d_m)$, and a polynomial function $e(\vec{x})$ is developed such that it yields estimates of the probability of relevance $P(R \mid \vec{x}(f_l, d_m))$ with minimum square errors. We give experimental results for the application of this approach to documents with weighted indexing as well as to documents with complex representations. In contrast to other probabilistic ... ery complex ... ed relevance ... and it needs

... least squares ... d Indexing— ... Search and

... al functions, ... ility ranking

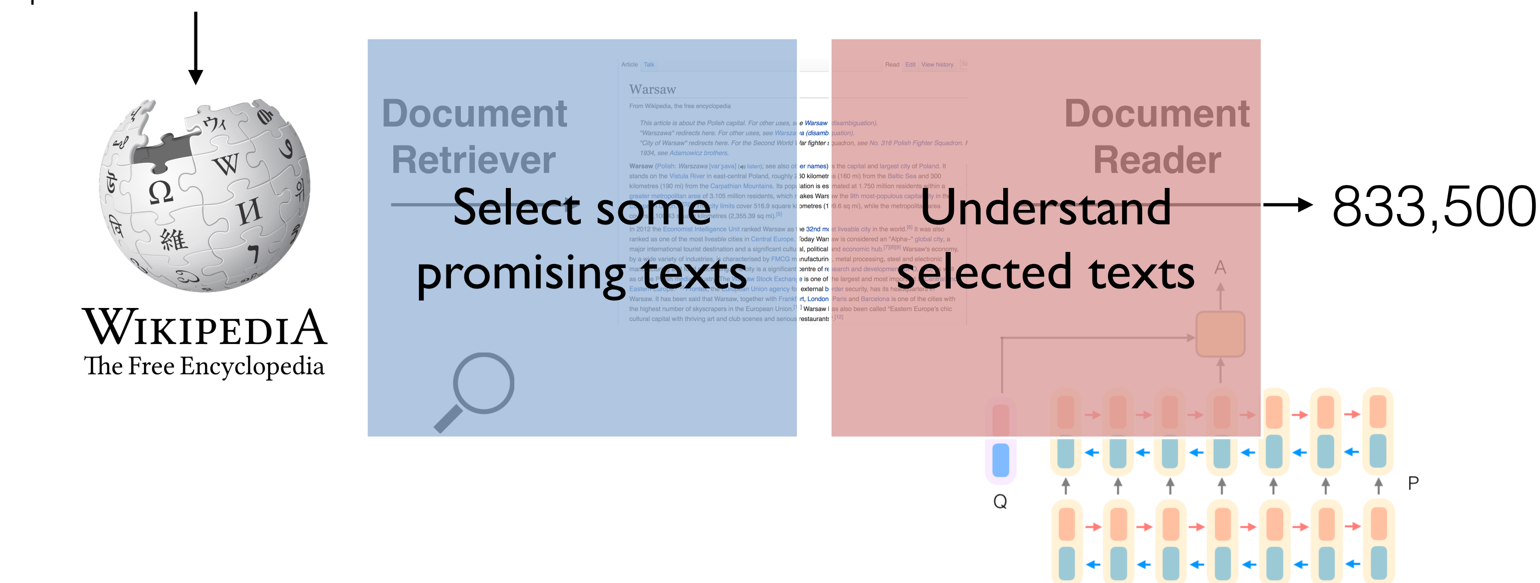


... methods. ... (explicitly ... rtain concept ... o perfect re... e all relevant documents ahead of the first nonrelevant one), optimum retrieval is defined with respect to certain realistic restrictions concerning the retrieval process. Furthermore, some evaluation criteria must be given. Then optimum retrieval is the best retrieval (in terms of the evaluation criteria) that can be achieved while following

Reading Wikipedia to Answer Open-Domain Questions

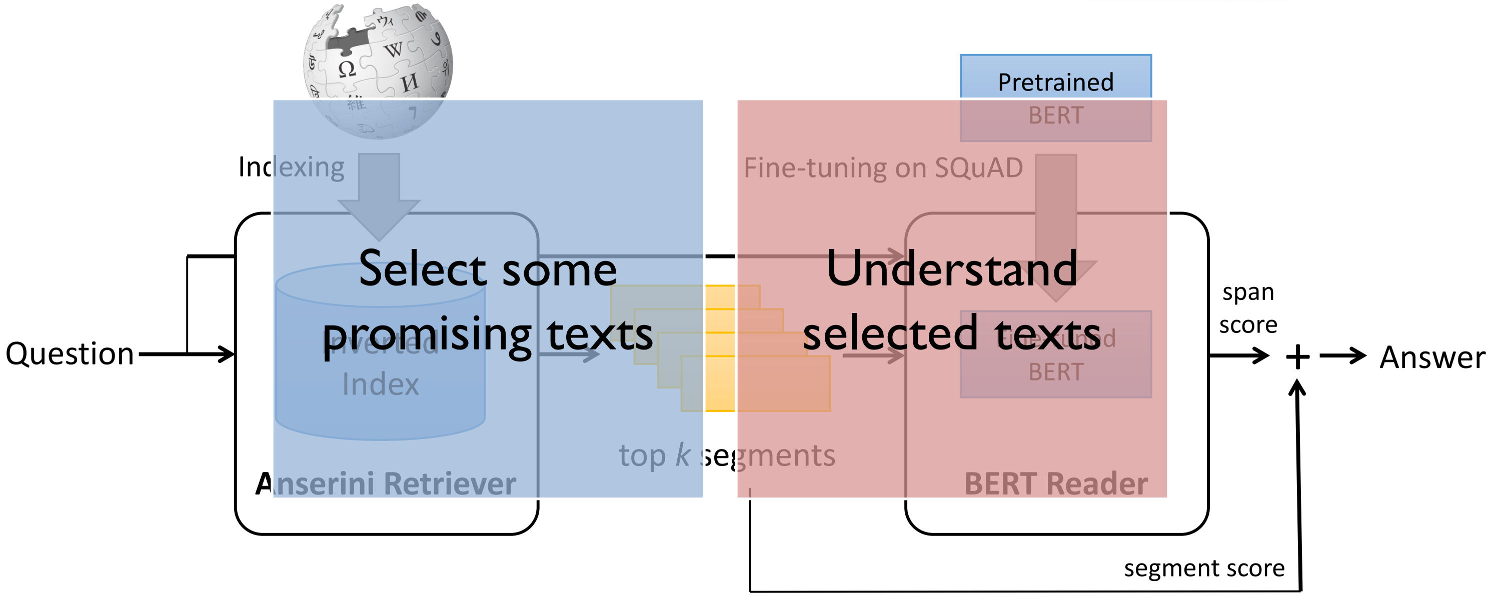
Chen et al. (ACL 2017)

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



BERTserini

question answering



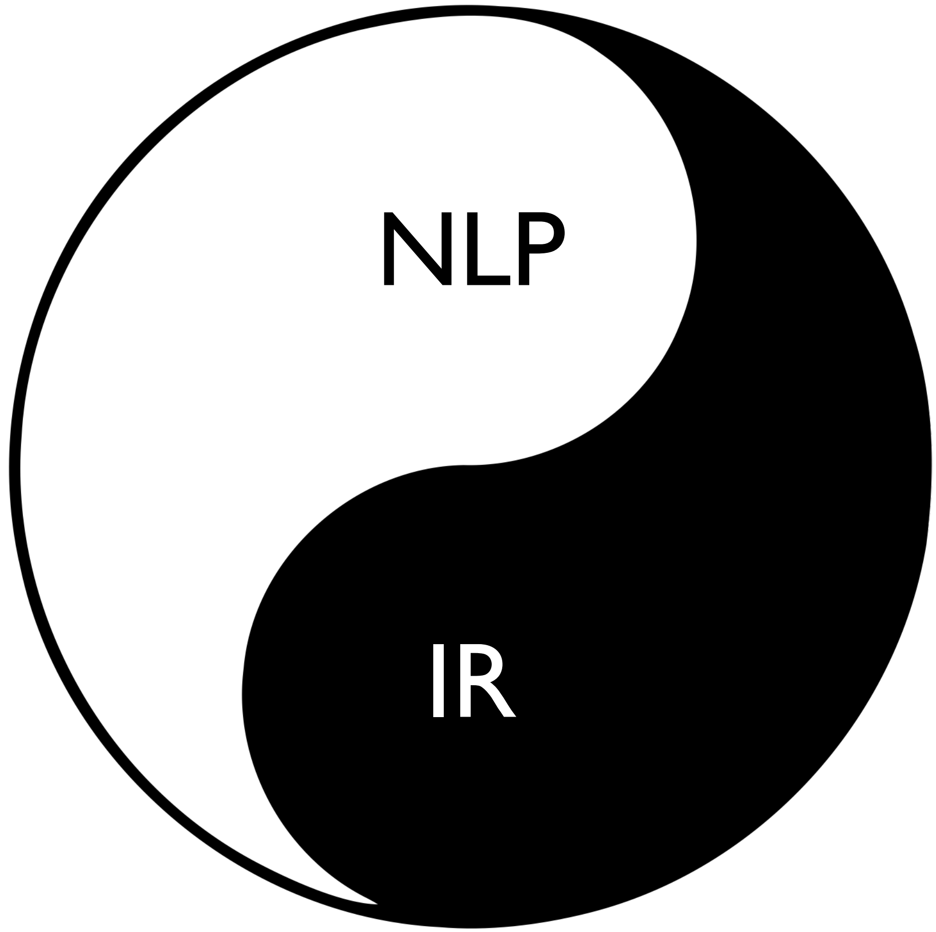
Yang et al. End-to-End Open-Domain Question Answering with BERTserini. *NAACL 2019 demo.*

Working hypothesis, *revised*:

solving the information access problem requires the synthesis of NLP and IR

BERT for question answering?
BERT for *ad hoc* retrieval?

Wow!



Together at last!

Loose Ends…

What is it about muppets?

Back to understanding…

Two steps at once?

It’s an exciting time to do research!

Loose Ends…

What is it about muppets?

Back to understanding…

Two steps at once?

Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

Select some promising texts

Understand selected texts

Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

Select some promising texts

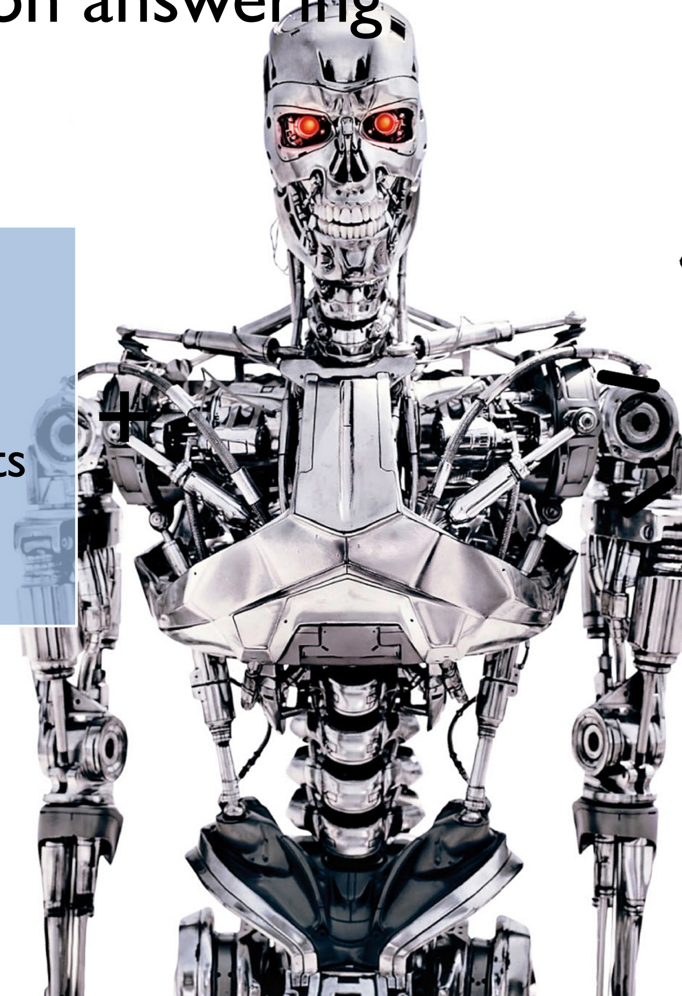
+

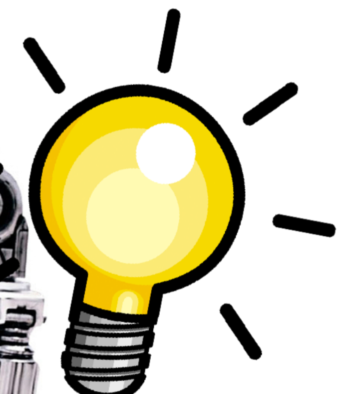
What is it about BERT?

Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

Select some promising texts





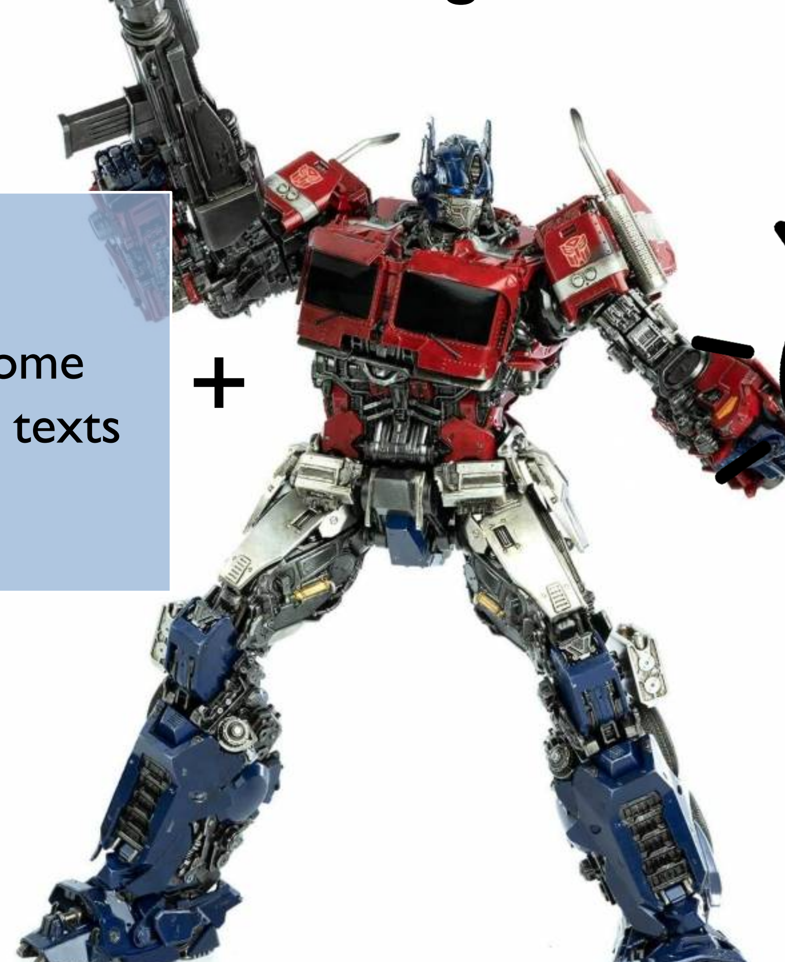
Ranking with T5 is even better!
(See recent results from the TREC-COVID challenge)

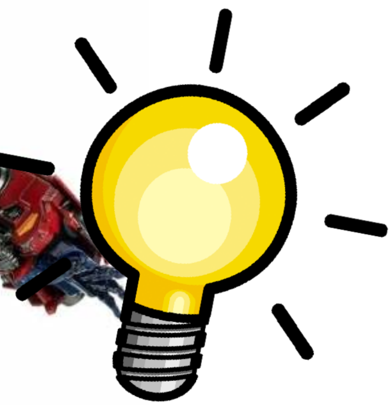
Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

Select some promising texts

+





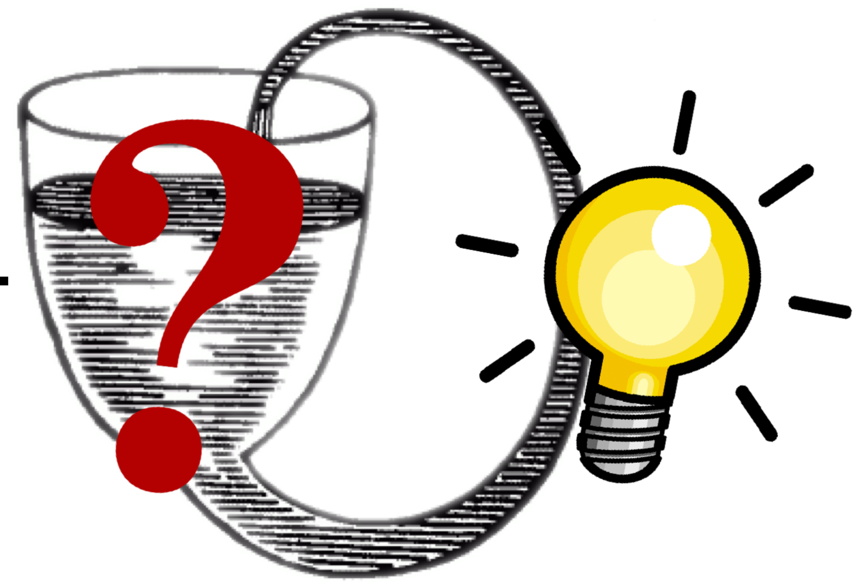
So it's about transformers?

Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

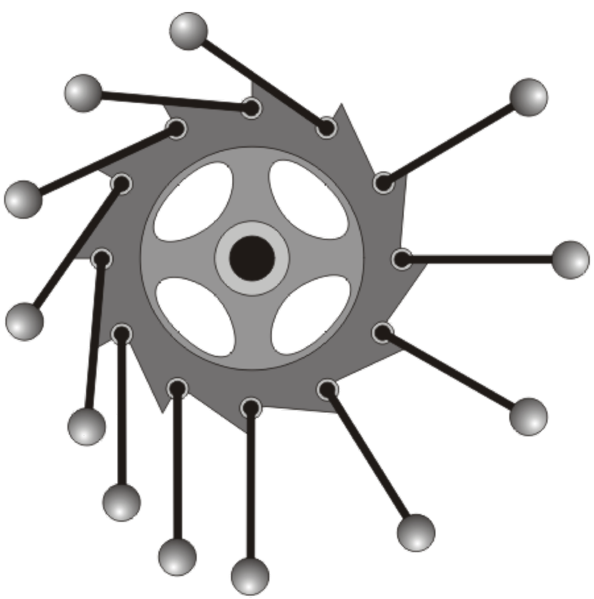
Select some promising texts

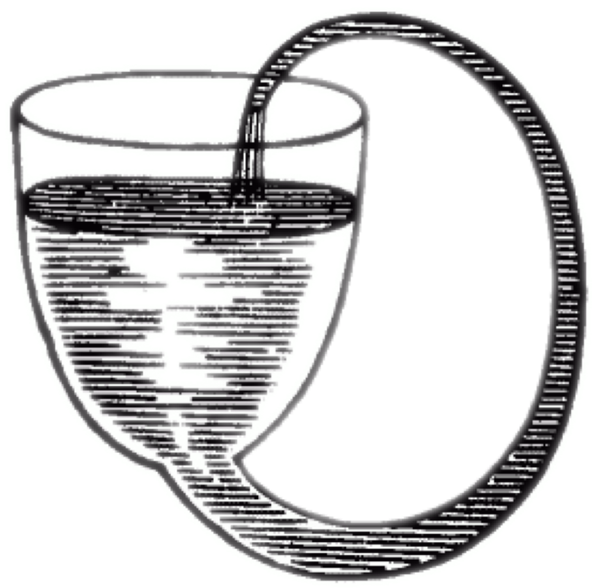
+

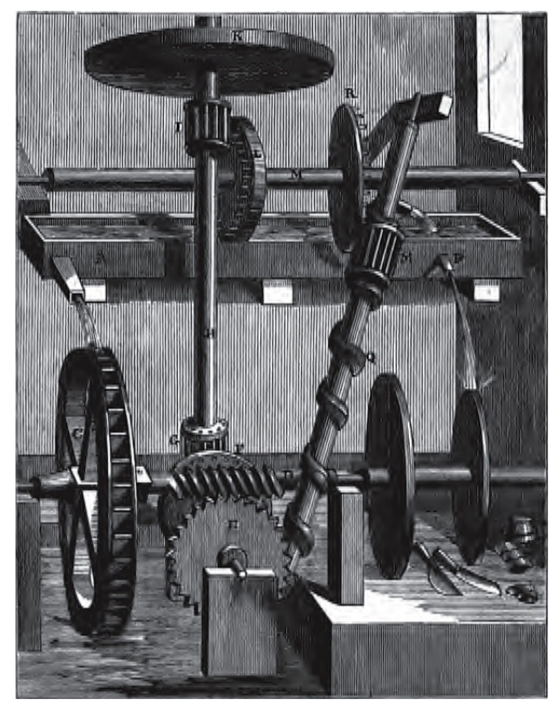


Perpetual Motion Machine

Perpetual motion is the motion of bodies that continues forever.
A perpetual motion machine is a hypothetical machine that can do work infinitely without an energy source. This kind of machine is impossible, as it would violate the first or second law of thermodynamics.

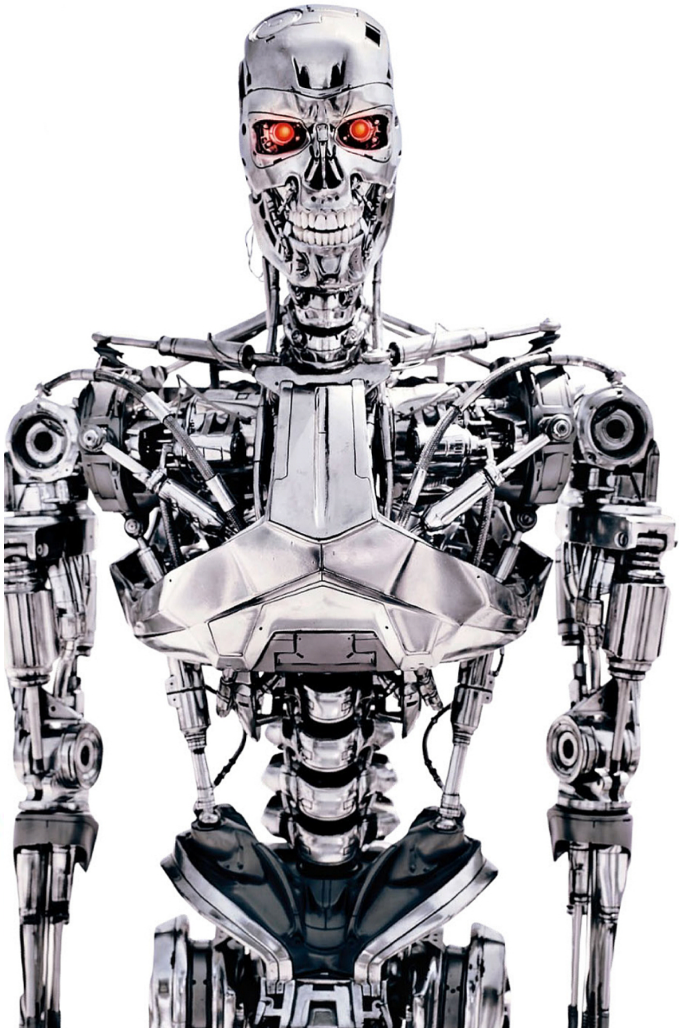






Source: Wikipedia

NLP's Perpetual Motion Machines

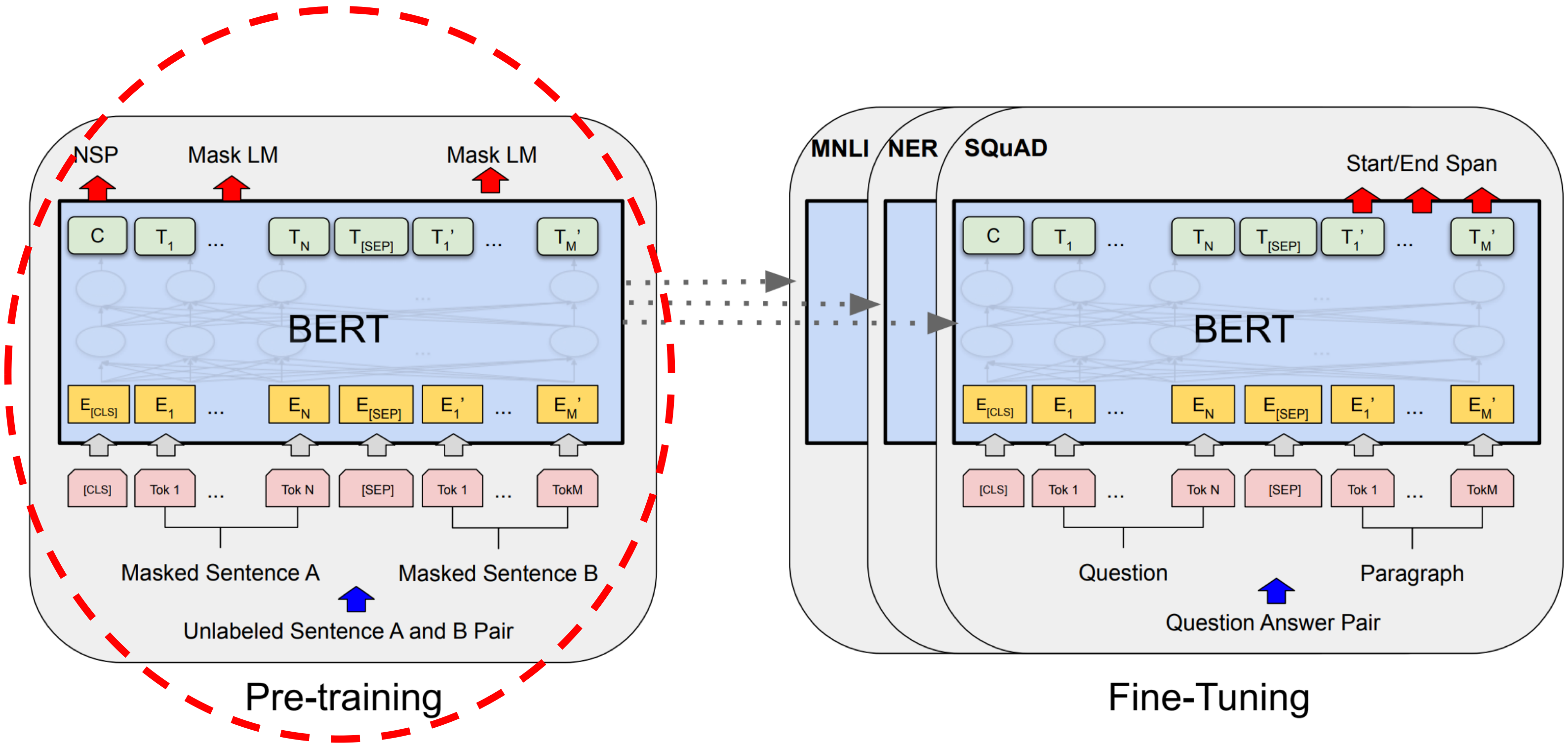


The secret ingredient?

Source: Food Network

The secret ingredient?

Self supervision!



Transformers w/ MLM was the successful example!

No doubt, the secret ingredient can be applied in other ways!

Where do we go from here? What's next?

I don't know… but I find this very exciting!

(Maybe Luke has the answers?)

Loose Ends…

What is it about muppets?

Back to understanding…

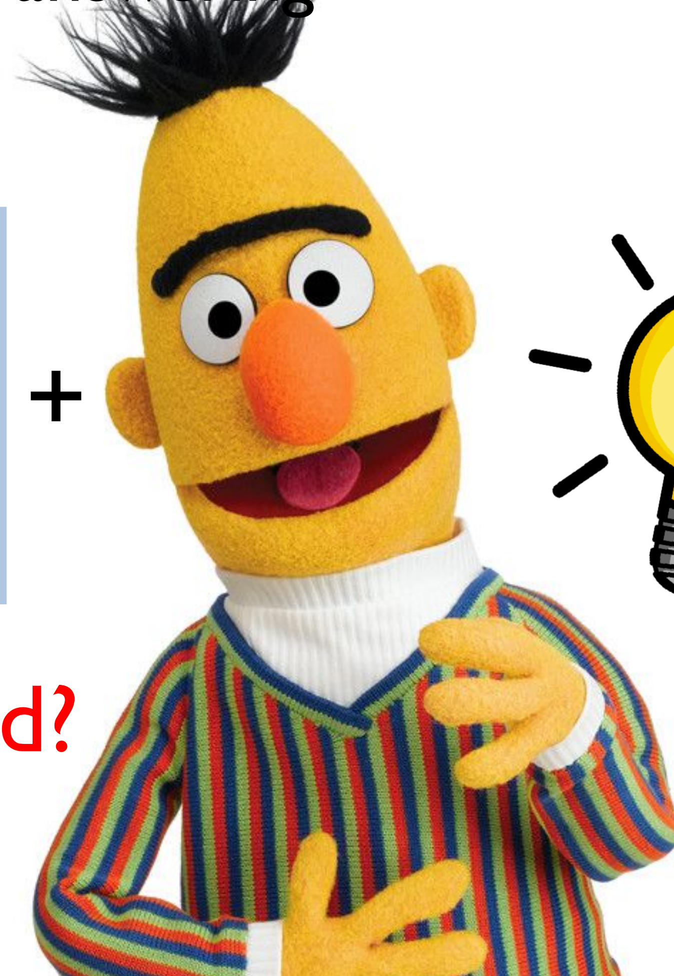
Two steps at once?

Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

Select some promising texts

+



Does BERT understand?

NO

(But I *don't* think the question is interesting)

Turing, Octopi, Chinese rooms…


My career-long quest…
Connecting users with relevant information
Understanding is what understanding does!

What does "understanding" mean?

For this talk, I'll treat it like pornography.

I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description ["hard-core pornography"], and perhaps I could never succeed in intelligibly doing so. But ***I know it when I see it***…

U.S. Supreme Court Justice Potter Stewart
in *Jacobellis v. Ohio* (1964)

counting the frequency of terms
identifying named entities
syntactic parsing
semantic role labeling

Increasing "understanding"

Where does BERT belong?

My Complaint about NLP

Most of NLP is focused on component techniques:

POS tagging, NER, relation extraction, parsing, SRL
paraphrase detection, sentiment analysis, etc.

There aren't many *extrinsic* tasks in NLP!

Information access is one of them
(machine translation is the other big one)

The quest for "understanding"?
Understanding for what?

Understanding is what understanding does!

An Operational Perspective

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.*

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway

The model *appears* to understand the text.

Jia and Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. *EMNLP 2017.*

An Operational Perspective

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

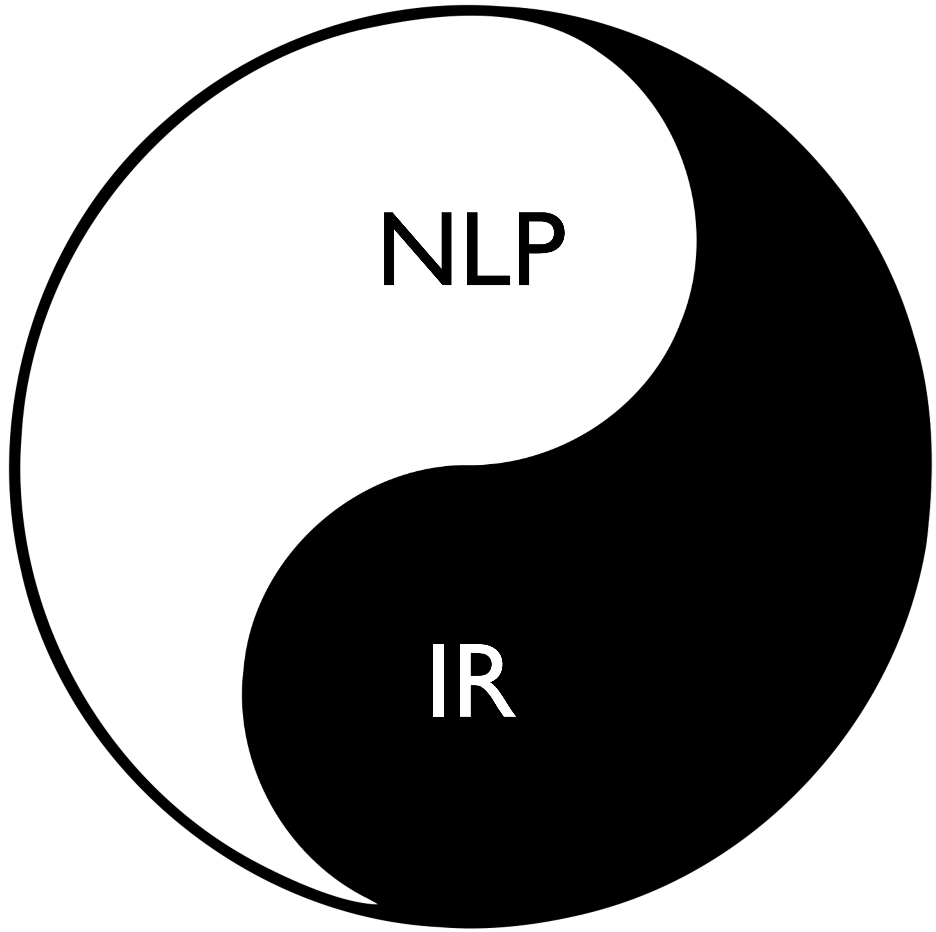
Clearly, the model is not understanding.

See, wasn't that easy?

Jia and Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. *EMNLP 2017.*

IR makes NLP useful.

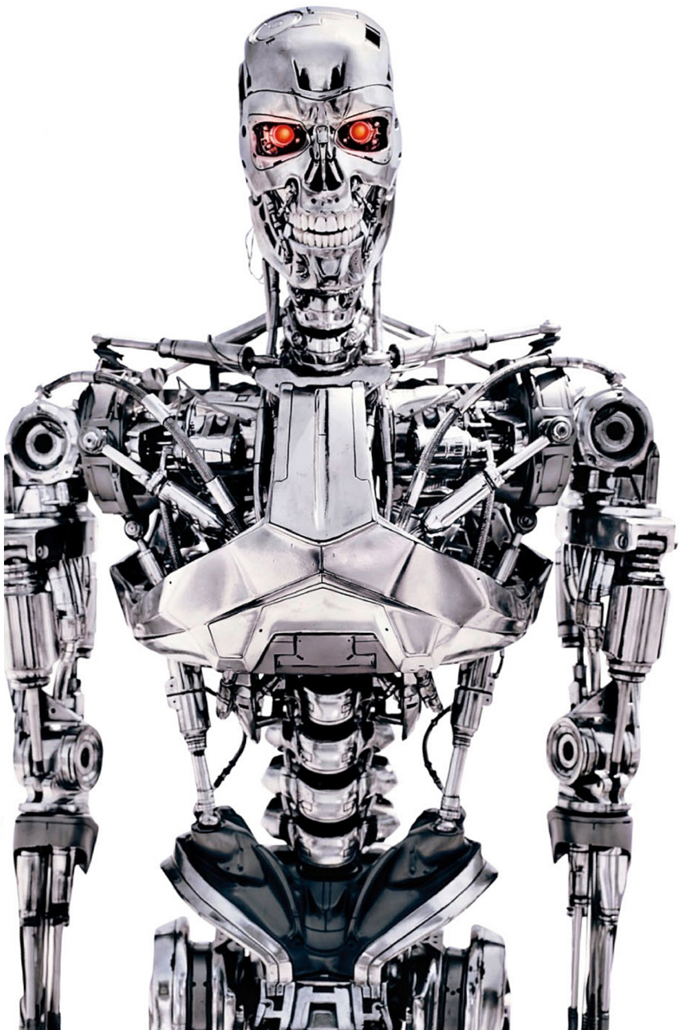
(Gives NLP something to do!)



NLP makes IR interesting.

(Moves beyond counting features!)

It works, but why?



It works, but why?

BERT works better with natural language input.
Removing stopwords *decreases* effectiveness!

Dai and Callan. Deeper Text Understanding for IR with Contextual Neural Language Modeling. *SIGIR 2019*.

"… certain attention heads correspond well to linguistic notions of syntax and coreference…"

Clark et al. What Does BERT Look At? An Analysis of BERT's Attention. *BlackBoxNLP 2019*.

"[BERT] … represents the steps of the traditional NLP pipeline in an interpretable and localizable way…"

Tenney et al. BERT Rediscovers the Classical NLP Pipeline. *ACL 2019*.

It works, but why?

Surely, there is some "understanding" going on here?

Let's figure it out!

Information Access

The challenge of scale

The challenge of understanding

The same?

More data, bigger model!

Even if… it's still interesting!

GPT-3

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

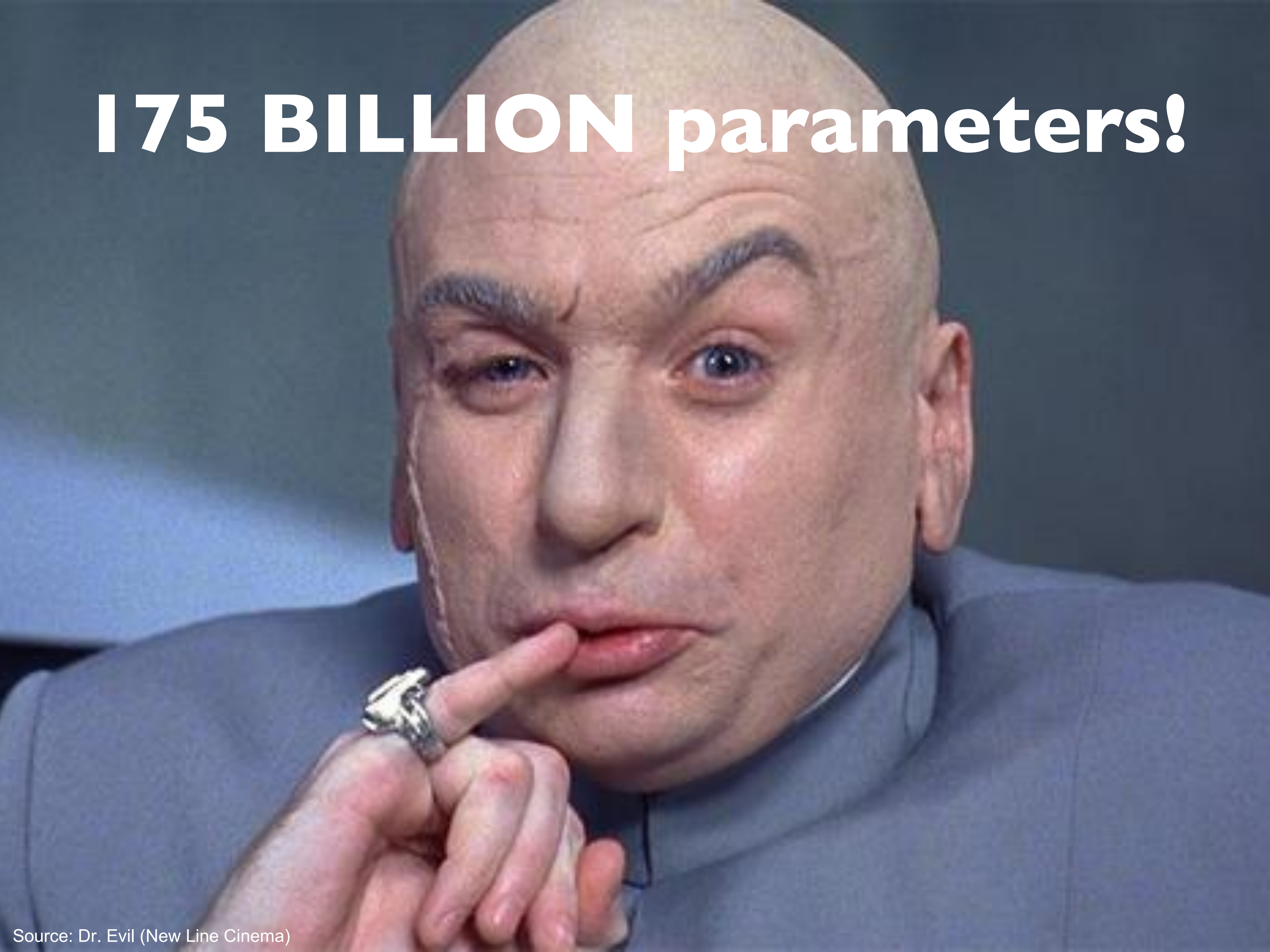
Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as



Source: Dr. Evil (New Line Cinema)

175 BILLION parameters!

GPT-3

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

arXiv:2005.14165v3 [cs.CL] 5 Jun 2020

I look at GPT-3 and I'm *not* depressed.

We know brute force works!

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few ex[illegible] largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion [illegible] the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as

How can we be smarter?

I don't know, but the answer will be very exciting!

Loose Ends…

What is it about muppets?

Back to understanding…

Two steps at once?

Information Access in Two Steps

document (*ad hoc*) retrieval
question answering

Select some promising texts

Understand selected texts

Information Access in a One Step?

document (*ad hoc*) retrieval
question answering

(dense vector retrieval stuff…)

Lee et al. Latent Retrieval for Weakly Supervised Open Domain QA. *ACL 2019*.
Reimers and Gurevych. Sentence-BERT. *EMNLP 2019*.
Humeau et al. Poly-encoders. *ICLR 2020*.

Loose Ends…

What is it about muppets?

Back to understanding…

Two steps at once?

It’s an exciting time to do research!

Questions?