Natural Language Processing for Information Retrieval: Challenges and Opportunities

ChengXiang Zhai

Department of Computer Science University of Illinois at Urbana-Champaign <u>http://www.cs.uiuc.edu/homes/czhai</u>



What is Information Retrieval (IR)?

- Salton's definition (Salton 68): "information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information"
 - Information: mostly text, but can be anything (e.g., multimedia)
 - Retrieval:
 - Narrow sense: search/querying
 - Broad sense: information access; information analysis

In more general terms

- Help people manage and make use of all kinds of information

Users are always an important factor!



NLP as Foundation of IR



IR researchers have been concerned about NLP since day one...



Luhn's idea (1958): automatic indexing based on statistical analysis of text



Hans Peter Luhn (IBM)

"It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements. " (Luhn 58)

LUHN, H.P., 'A statistical approach to mechanised encoding and searching of library information', *IBM Journal of Research and Development*, **1**, 309-317 (1957). LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, **2**, 159-165 (1958).



The notion of "resolving power of a word"



Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adaped from Schultz⁴⁴page 120)

The Database and Information Systems Laboratory at the University of Illinois at Ubana Charangen The University of Illinois at Ubana Charangen

Automatic abstracting algorithm [Luhn 58]



Sentence

Portion of sentence bracketed by and including significant words not more than four non-significant words apart. If eligible, the whole sentence is cited.

The idea can be adapted for query-specific summarization

"In many instances condensations of documents are made emphasizing the relationship of the information in the document to a special interest or field of investigation. In such cases sentences could be weighted by assigning a premium value to a predetermined class of words."

Figure 2 Computation of significance factor. The square of the number of bracketed significant words (4) divided by the total number of bracketed words (7) = 2.3.



Cleverdon's Cranfield Project (1957-1966)



Established rigorous evaluation methodology Introduced precision & recall Compared different linguistic units for indexing

Cyril Cleverdon Cranfield Inst. of Tech, UK





Indexing and Abstracting by Association

Doyle, Lauren B, American Documentation (pre-1986); Oct 1962;



FIG. 1. An Association Map Based on Pearson Correlation Coefficients.

The Database and Information Systems Laboratory at the luxership of Illinois at Usana Charagely Large Scale (Information Monogeneri And many attempts have been made on improving IR with NLP techniques since then...

However, today's search engines don't use much NLP!



Sometimes, they appear to "understand" natural language

Query: "NLP & CC 2012"

Ad related to NLP & CC

NLP Certification | | www.nlpcoaching.com/ Fastest NLP Certificatio

"NLP&CC 2012"征 tcci.ccf.org.cn/conferen NLP&CC 2012. 2012年 Oct.31日~Nov.5,2012, 1

<u>中国计算机学会中文</u> tcci.ccf.org.cn/.../**2012**/p Natural language proce most popular research ;

NLP&CC 2012投稿 ② tcci.ccf.org.cn/.../2012/p

2012年6月15日. 2012年 长一周. NLP&CC 2012;

<u>第一届自然语言处理</u> www.meeting.edu.cn→i 第一届自然语言处理与⊑ 创新与应用"为主题,重

The Database and Information Systems Laboratory

1,390,000 RESULTS Any time

www.falconevw.com Visit Falcone Volkswagen & See

<u>中国计算机学会中文信息技</u> tcci.ccf.org.cn/conference/2012/in ◆ 2012年6月15日, NLP&CC 20 ◆ 2012年6月8日, NLP&CC 201;

<u>中国计算机学会中文信息技</u> tcci.ccf.org.cn/conference/2012/pr NLP&CC 2012注册: 自助餐: 自助 第七讲: NLP&CC 2012注册: ◆ W

<u>第一届自然语言处理与中文</u> mall.cnki.net/magazine/Article/JS、 第一届自然语言处理与中文计算会 期杂志在线阅读、文章下载。<正

<u>第一届自然语言处理与中文</u> www.cascs.org/thread-520-1-1.ht NLP&CC 2012以"互联网创新与应 言处理与中文计算,包括Web挖掘

1884-CC Morgan MS62 G

TIMAN

中国计算机学会中文信息技术专业委员会-2012年会介绍 ❷ NLP&CC 2012简介: 自然语言处理与中文计算会议(NLP&CC)是由中国计算机学会(CC

F)主办的CCF中文信息技术专业委员会年度学术会议。NLP&CC专注于自然语言处理及... tcci.ccf.org.cn/conference/**2012 2**012-10-11 - <u>百度快照</u>

NLP&CC2012会议注册及成果展示通知-中国计算机学会信息网 ∅

NLP&CC2012会议注册及成果展示通知 自然语言处理与中文计算会议(以下简称NLP&CC) 是由中国计算机学会(CCF)主办的CCF中文信息技术专业委员会年度学术会议。NLP&CC... www.ccf.org.cn/sites/ ... 689087728707 2012-10-12 - <u>百度快照</u>

第一届自然语言处理与中文计算会议(NLP&CC 2012)) 文章类型:学术会议 发布时间... 开 始日期: 2012-10-31 结束日期: 2012-11-05 所在国家: 中华人民共和... www.sjzu.edu.cn/news.asp?code=28996 2012-9-28 - 百度快照

第一届自然语言处理与中文计算会议(NLP&CC2012)征文通知-【维... 由中国计算机学会(CCF)主办的CCF中文信息技术专业委员会年度学术会议一一自然语言处 理与中文计算会议(以下简称NLP&CC)将于2012年10月31日--11月5日在北京召开。 www.cqvip.com/QK/... 1248743.html 2012-10-4 - 百度快照

Chinese Weibo Sentiment Analysis Evaluation at NLP&CC 2012

This post summarize our BITSTAR stentiment analysis system paticipating in the Chinese W eibo Sentiment Analysis Evaluation at NLP&CC 2012. We propose a ... www.bitwjg.org/2012/09/17/chinese-w ... 2012-9-23 - 百度快照

第一届自然语言处理与中文计算会议(NLP&CC 2012) 会议网站: http://tcci.ccf.org.cn/conference/2012/ 会议背景介绍: 自然语言处理与中文计算



Query: "NLP & CC 2012 schedule"

Web	Ad related to NLP & CC 2012 Schedule ()
Images	NLP Certification NIpCoaching.com www.nlpcoaching.com/
Maps	Fastest NLP Certification possible! Get NLP Certified in only 7 days.
Videos	Cornell NLP Seminar, Spring 2012 - Cornell NLP group - Confluence
News	https://confluence.cornell.edu//NLP/Cornell+NLP+Seminar,+Spring Apr 28, 2012 – This wiki page.
Shopping	https://confluence.cornell.edu/display/NLP/Cornell+NLP+Seminar% 2C+Spring+2012, holds the schedule for the NLP seminar
More	
	Cornell NLP Seminar, Fall 2012 - Cornell NLP group - Confluence
	https://confluence.cornell.edu//NLP/Cornell+NLP+Seminar,+Fall+2

Champaign, IL Change location

Show search tools

Home - Cornell NLP group - Confluence

holds the schedule for the NLP seminar for ...

2 days ago - This wiki page,

https://confluence.cornell.edu/display/NLP/Home Jul 30, 2012 – The calendar maintains info about NLP-related talks on campus for the ... Claire Cardie are nominated for best paper award at SIGDial 2012 ...

https://confluence.cornell.edu/display/NLP/Cornell+NLP+Seminar%2C+Fall+2012,

natural language processing blog: Somehow I totally missed NIPS ...





How does a typical search engine work? Bag of Terms Representation





A Typical Ranking Function



Feedback in IR



The Database and Information Systems Laboratory



Search Engines Generally Do Little NLP

- Bag of words representation remains the pillar of modern IR
- Simple lexical processing: stop words removal, stemming, word segmentation
- Limited uses of phrases

+ ...

Basic Technique = Keyword Matching

- + statistical weighting of terms
- + leveraging clickthroughs (feedback)

NLP <= Lexical Analysis (?)



IR researchers don't talk much about NLP today either

Assumed Conclusion: NLP isn't useful for IR...



Questions

- If logically NLP is the foundation of IR, why hasn't NLP made a significant impact on IR?
- Is there any way to improve IR significantly with NLP techniques?
- What does the future of NLP for IR look like?



Rest of the Talk

- Attempts on applying NLP to IR
- Why hasn't it be successful?
- The future of NLP for IR



NLP for IR 1: Beyond bag-of-words Representation

Motivation: single words have many problems

Different words, same meaning: car vs. vehicle

Same words, different meaning: Venetian Blinds vs. blind venetians

Different perspectives on single concept: "The accident" vs. "the unfortunate incident"

Different meanings for the same words in different domains: "sharp" can mean "pain intensity" or "the quality of a cutting tool"

[Smeaton's ESSIR'95 tutorial]



Many different phrases explored

- Statistical phrases [Fagan 88]
 - Phrases are frequent n-grams
- Linguistic phrases [Fagan 88, Zhai & Evans 96]
 - Phrases are obtained from parsing
- Lexical atoms [Zhai et al. 95; Zhai 97]
 - "Sticky phrases"/non-compositional phrases (e.g., "hot dog", "blue chip")
- Head-modifier pairs [Strzalkowski & Vauthey 95, Zhai 97]

"fast algorithm for parsing context-free languages"

➔{"fast algorithm", "parsing algorithm", "parsing language", "context-free language"}



Phrase Indexing: Results

- Mostly mixed results
 - Some reported insignificant improvement over single word baseline
 - Others reported degradation of retrieval accuracy
- While on average, using phrases may help, it doesn't help all queries
- Even when adding phrases helps, adding "too many" phrases can hurt the performance
- Mixing phrases with single words is generally necessary to improve robustness



Sample Phrase Indexing Results [Zhai 97]

Experiments	Recall (Ret-Rel)	Init Prec	Avg Prec			
WD-SET	0.56(597)	0.4546	0.2208			
WD-HM-SET	0.60(638)	0.5162	0.2402			
inc over WD-SET	7%	14%	9%			
WD-NP-SET	0.58(613)	0.5373	0.2564			
inc over WD-SET	4%	18%	16%			
WD-HM-NP-SET	0.63(666)	0.4747	0.2285			
inc over WD-SET	13%	4%	3%			
Total relevant documents: 1064						

Table 1: Effects of Phrases with feedback and TREC-5 topics

Too many phrases hurt performance!



NLP for IR 2: Sense Disambiguation

Motivation

- Terms are often ambiguous, causing mismatches
- What about using term disambiguation?

Many studies

- Krovetz and Croft 1992
- Voorhees 1993
- Sanderson 1994
- Schultz and Pedersen 1995
- Stokoe et al. 2003



Disambiguation Results: Non-Promising

- Manual sense disambiguation [Korvetz & Croft 92]
 - Very little improvement (<=2% improvement)
 - Possibly degrade performance
 - Explanation: coordination of terms; skewed distribution of senses
- Automatic sense disambiguation based on WordNet [Voorhees 93]
 - No improvement

"pseudo sense" experiments [Sanderson 94]

IR performance is very sensitive to erroneous disambiguation ... Only when it gets to 90% accuracy it is as good as no disambiguation... Beyond that, it yields improvement, but only when the query is short



Disambiguation Results: More Promising

- Corpus-based senses [Schultz & Pedersen 92]
 - Senses are acquired by clustering word context
 - Multiple senses are assigned to combat uncertainty
- Semcor 1.6 + careful weighting [Stokoe et al. 03]



Figure 3. Plots precision for 11 recall points for the Term Based, Stem Based, Sense Based (T), and Sense Based (S) retrieval runs



NLP for IR 3: Deeper Semantic Representation: FERRET [Mauldin 91]

- using knowledge representation to represent text
- works for a very small data set in astronomy domain
- but, doesn't scale up, possibly not outperforming stronger baseline_____





Rest of the Talk

- Attempts on applying NLP to IR
- Why hasn't it be successful?
- The future of NLP for IR



Explanation 1: The Power of Bag of Words Representation

- Retrieval problem is mostly a simple language processing task
- "Matching" is sufficiently useful for finding relevant documents
- Ideal query hypothesis: given any subset of documents that we assume a user is interested in, there exists a query that would produce nearideal ranking
- Finding an ideal query doesn't necessarily need deep NLP



Keyword matching may answer questions!



how to install hardwood floor?

177,000,000 RESULTS

Tips: Current page contains all results En English Search | Chinese Only

How to Install Solid Wood Floors | eHow.com

www.ehow.com/how_5645653_install-solid-wood-floors.html How to Install Solid Wood Floors in Basements The problem with trying to install a hardwood floor in a basement is that basement floors usually are concrete slabs, and ...

How to Install a Wood Floor | eHow.com

www.ehow.com/how_2325521_install-wood-floor.html But before you begin installing a hardwood floor, you need to... How to Install Real Wood Flooring Solid wood flooring is one of the single most durable and beautiful flooring ...

How to Install a Hardwood Floor | HomeTips

www.hometips.com/diy-how-to/installing-hardwood-floors.html Expert advice on hardwood floor installation from start to finish, including tools and materials, preparation, layout, cutting, and fastening ... Expert advice on hardwood floor ...

How to Install Hardwood Floors- Do it Yourself Hardwood Floors www.diynetwork.com/topics/hardwood-floors/index.html Easy-to-follow, step-by-step instructions show DIYers how to install a floating wood plank ... Marc Bartolomeo shows how to install a beautiful engineered hardwood floor.



Explanation 2: NLP wasn't used to solve a big pain

Different words, same meaning: car vs. vehicle⁴

Same words, different meaning: Venetian Blinds vs. blind venetians

Different perspectives on single concept:

"The accident" vs. "the unfortunate incident"

Different meanings for the same words in different domains: "sharp" can mean "pain intensity" or "the quality of a cutting tool"

Some times domain restriction solves the problem naturally. Other words in the query help providing disambiguation.



How likely does this happen?

feedback & expansion

can take care of this

Explanation 3: Lack of consideration of robustness

- Standard IR models are optimized for bag of terms representation
- When incorporating phrases, we no longer have optimal term weighting
 - e.g., how to optimize phrase weighting when single words are also used for indexing?
- Need to tolerate NLP errors



Explanation 4: Workaround possible



Queries



Example of NLP for tail queries: sense clarification [Kotov & Zhai 11]

- Uses global analysis for sense identification:
 - + does not rely on retrieval results (can be used for **difficult queries**)
 - +identifies collection-specific senses and avoids the coverage problem
 - + identifies both majority and minority senses
 + domain independent
- Presents concise representations of senses to the users:

+eliminates the cognitive burden of scanning the results

Allows the users to make the final disambiguation choice:

+leverages user intelligence to make the best choice



Query ambiguity





Query ambiguity





Sense feedback improves retrieval accuracy on difficult topics

		KL	KL-PF	SF
AP88-89	MAP	0.0346	0.0744	0.0876
	P@10	0.0824	0.1412	0.2031
ROBUST04	MAP	0.04	0.067	<u>0.073</u>
	P@10	0.1527	0.1554	0.2608
AQUAINT	MAP	0.0473	0.0371	<u>0.0888</u>
	P@10	0.1188	0.0813	0.2375

• Sense feedback outperforms PRF in terms of MAP and particularly in terms of Pr@10 (**boldface** = statistically significant (p<.05) w.r.t. KL; <u>underline</u> = w.r.t. to KL-PF)



Rest of the Talk

- Attempts on applying NLP to IR
- Why hasn't it be successful?
- The future of NLP for IR



Future of NLP for IR: Challenges

- Grand Challenge: How can we leverage imperfect NLP to create definite value for IR?
- Possible Strategies
 - Create add-on value: supplement rather than replace existing IR techniques
 - Integrate NLP into a retrieval model (minimize "disruption")
 - Multi-resolution representation
 - Include users into the loop



NLP for IR Opportunity 1: long-tail queries

NLP for query understanding in context

- Query segmentation
- Query parsing
- Query interpretation
- Do all these in the context of search session and user interaction history

NLP for document browsing

- When querying fails, browsing helps
- How to create a multiresolution topic map?
- NLP for interactive search
 - How to generate clarification questions?



NLP for IR Opportunity 2: Beyond Topical Relevance

- Traditional IR work has focused on exclusively topical relevance
- Real users care about other dimensions of relevance as well
 - Sentiment/Opinion retrieval: find positive opinions about X
 - Readability: find documents with readability level of elementary school students
 - Trustworthiness
 - Genre







Towards an intelligent knowledge service system



Summary

- NLP is the foundation of IR, but keyword matching is quite powerful
- NLP for IR hasn't been so successful because of the focus on document retrieval (narrow sense of IR)
- Many more opportunities in applying NLP to IR in the future
 - Need to supplement, rather than replace existing IR techniques
 - Aim at more intelligent, interactive knowledge service system



Thank You!

Questions/Comments?



