

# To Click or not to Click?

## The Role of Contextualized and User-Centric Web Snippets

Nikos Zotos – Patras University  
 Paraskevi Tzekou – Patras University  
 George Tsatsaronis – Athens University  
 Lefteris Kozanidis – Patras University  
 Sofia Stamou – Patras University  
 Iraklis Varlamis – Athens University

9/8/2007

SIGIR-2007-FOCUS

1

## Lost in Search Results



### RESULTS

- Too many for a query
- Convey little information about their relevance to the query

**Challenge:** How do users choose the URLs to click on?

9/8/2007

SIGIR-2007-FOCUS

2

## Clicking Decisions

- URLs might help
- Title represents the page contents
- Text fragments give a glimpse to the page's content

[Statistical NLP / corpus-based computational linguistics resources](#)  
 An annotated list of resources in this field and the allied discipline of **statistical natural language processing**. Corpora, tools, literature and other ...  
[www.nlp.stanford.edu/links/statnlp.html](#) - 16k - [Cached](#) - [Similar pages](#)

[Foundations of Statistical Natural Language Processing](#)  
 Companion web site for the book, published by MIT Press, June 1999.  
[nlp.stanford.edu/fnlp/](#) - 7k - [Cached](#) - [Similar pages](#)

[Foundations of Statistical Natural Language Processing](#)  
 Promotional Web Site for the Book, published by MIT Press, May 1999.  
[nlp.stanford.edu/fnlp/promo/](#) - 5k - [Cached](#) - [Similar pages](#)

[Statistical Natural Language Processing: Models and Methods \(CS775\)](#)  
 Statistical approaches have revolutionized the way NLP is done. Furthermore, some of these approaches can be employed in other applications. ...  
[www.cs.cornell.edu/courses/cs775/2001sp/default.html](#) - 21k - [Cached](#) - [Similar pages](#)

[Ling 684.02 Class, WI 2002](#)  
 Natural Language Processing. ... Thursday 1/10 Statistical (HTML for Explorer only -- sorry) ...  
[brew795M/](#) - 15k - [Cached](#) - [Similar pages](#)

[rocessing - Wikipedia, the free encyclopedia](#)  
 e processing uses stochastic, probabilistic and statistical of the difficulties discussed above, ...  
[al\\_language\\_processing](#) - 42k - 6 Jul 2007 - [Cached](#) - [Similar pages](#)

**Snippet:** A set of contiguous text extracted from a page using statistical methods

## Snippet Selection

### Statistical Natural Language Processing

This web-based course in **statistical natural language processing** is meant to provide the basic material for a distance learning course, although some local ...

[www.msi.vxu.se/users/nivre/teaching/statnlp/](#) - 3k - [Cached](#) - [Similar pages](#)

**vaxjo University**  
 School of Mathematics and Systems Engineering  
 Vaxjo University

### Web Course

The web-based course in statistical natural language processing is meant to provide the basic material for a distance learning course, although some local tutoring will normally be required. The course is built around three components:

- **Basic course:** An introductory reading course with lecture notes, slides, exercises, and pointers to the literature.
- **Projects:** A set of student projects, of varying size and complexity, meant to give the students practical experience and deepen their understanding of di methods in statistical NLP.
- **Tools and resources:** An inventory of available tools and resources for statistical NLP, to be used in carrying out the project work.

[Acknowledgements](#)

9/8/2007

SIGIR-2007-FOCUS

4

9/8/2007

SIGIR-2007-FOCUS

3

## The Role of Snippets

- Snippets do not really help the users decide on which URLs to click?

### WHY NOT?

- Not obviously related to the query intention
- Marginally informative of the pages' content
- Lack coherence, incomplete text

ANY IDEAS?

**Question:** Can we **improve** the contribution of snippets in the **decision** making process?

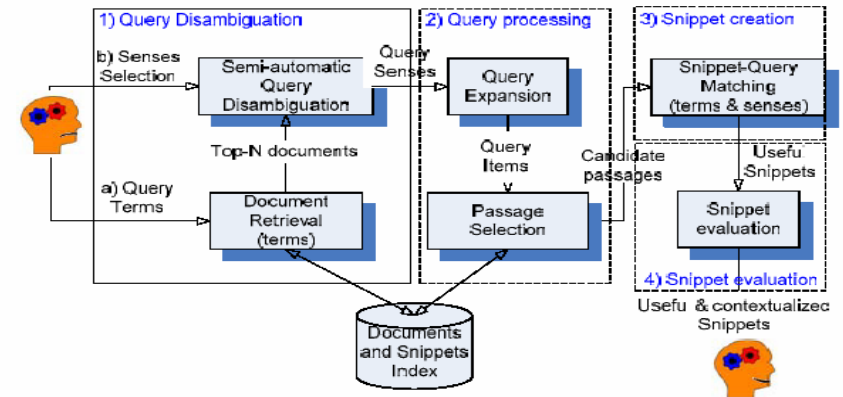
1. Select query-relevant snippets
2. Rely on text semantics
3. Examine coherence
4. Examine expressiveness
5. Ask users

9/8/2007

SIGIR-2007-FOCUS

## Motivation for our Study

**Design a sound model for snippet selection based on semantics**



9/8/2007

SIGIR-2007-FOCUS

6

## Our Contribution

### Hypothesis

- Semantic processing of both the query and the query-relevant pages will give better snippets
- Semantically-selected snippets will help the users make clicking decisions

Our contribution:  
Designed a method that selects **Expressive** and **Coherent** snippets by accounting for their **Usefulness** to the query intention

9/8/2007

SIGIR-2007-FOCUS

7

## Towards Useful Snippets

### Procedure

Given a query and a set of relevant pages

- (1) disambiguate the query intention
- (2) select candidate snippets based on their semantic similarity to the query
- (3) keep query-useful snippets
- (4) evaluate snippets' coherence and expressiveness
- (5) return best-matching snippet

9/8/2007

SIGIR-2007-FOCUS

8

## Identifying the Query Intention

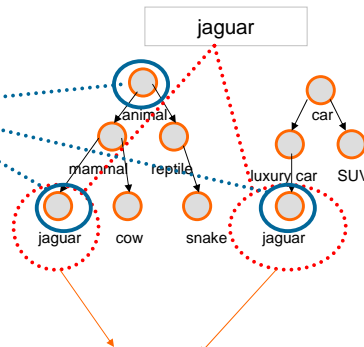
### Semi-automatic Query Sense Disambiguation:

- Map query keywords to WordNet
- Process the top N (N=20) returned pages and map their content terms to WordNet
- Take query matching-senses that have a semantic relation to the pages' content terms senses
  - Display them to the user
- Ask the user select the query sense that best describes the query intention



$S_1$ : Vehicle brand of which The 1932-1940 SS models except 4-cyl. Are classic cars.

$S_2$ : A large spotted feline of tropical America similar to the leopard, in some classifications considered a member of the genus *Felis*.



9/8/2007

SIGIR-2007-FOCUS

9

## Towards Query-Relevant Snippets

### Query and Page semantic processing:

- Expand query terms with their WordNet synonyms
- Find all appearances of query terms in the page's text
- Define window around query items

$$\text{Relevance}(q, p) = \frac{\sum_{j=1}^k q_r \cdot \text{Tf} / \text{IDF}(t_j, p)}{q_s \cdot \sum_{i=1}^n \text{Tf} / \text{IDF}(t_i, p)}$$

$k$ : terms in snippet  $p$  that relate to at least one term in  $q$   
 $n$ : total number of terms in  $p$   
 $q_r$ : number of  $q$  terms to which the snippet term  $t_j$  relates (query relevant terms)  
 $q_s$ : number of terms in the query (query size)  
 $\text{Tf}/\text{IDF}(t_x, p)$  importance of term  $t_x$  in  $p$  as determined by their cosine similarity VSM

- Compute **Semantic Correlation** between query and candidate snippet terms

$$\text{Similarity}(q_i, S_k) = \frac{2 * \text{depth}(\text{LCS}(i, k))}{\text{depth}(i) + \text{depth}(k)}$$

Based on the Wu and Palmer similarity metric

Least Common Subsumer

9/8/2007

SIGIR-2007-FOCUS

10

## Selecting Query-Useful Snippets

Query-snippet semantic correlation values weighted by the score of relation type ( $r$ ) quantifies the quality of the selected passage

$$\text{Quality}(S, q) = \frac{1}{n \times m} \sum_{j=1}^m \sum_{k=1}^n [\text{Similarity}(q_j, S_k) \cdot \text{RelationWeight}(r)]$$

Based on the findings of Song et al., 2004 (CICLing)

Synonymy: 1, Hyper/Hyponymy: 0.5, Mero/Holonymy: 0.4

Combining **Relevance** and **Quality** between snippet and query terms derives the snippet's **Usefulness** to the query intention

$$\text{Usefulness}(S, q) = \text{Relevance}(q, S) \cdot \text{Quality}(S, q)$$

Terminological Overlap  
Avg. semantic similarity

9/8/2007

SIGIR-2007-FOCUS

11

## Evaluating Selected Snippet

- Task I: Measuring Coherence
- Task II: Measuring Expressiveness

**Coherence**: indicates the degree of **in-snippet semantic** correlation and is useful in **selecting** the URLs to click on

**Expressiveness**: indicates the degree of semantic correlation between **snippet and remaining text** terms and is useful in **focusing** retrieval to useful text fragments

9/8/2007

SIGIR-2007-FOCUS

12

## Evaluating Selected Snippet (2)

**Semantic Coherence:** semantic similarity that snippet terms exhibit to each other, as determined in WordNet

$$\text{Coherence}(S_1) = \frac{1}{n} \sum_{i,j=1}^n \arg \max_{w_j} \text{similarity}(w_i, w_j)$$

Wu and Palmer metric

**Expressiveness:** semantic similarity that snippet and remaining text terms exhibit to each other

$$\text{Expressiveness}(S_1, (D - S_1)) = \text{Usefulness}(S_1, (D - S_1))$$

Product of: (i) Terminological Overlap (**Relevance**) between snippet terms and terms in the remaining document ( $D - S_1$ ) and (ii) Avg. **Semantic Correlation** between snippet and remaining document terms, weighted by the  $\text{Relation}(r)$  type.

## Experimental Study: Goals

- Examine performance of our semantically-driven snippet selection model

Compared the performance of our model to the performance of the **Alicante** statistical passage retrieval algorithm

- Examine influence of semantically-derived snippets on user decisions

Carried out a **blind user study with 15 participants** who were asked to make click decisions based on a number of different snippets offered for the same queries and pages

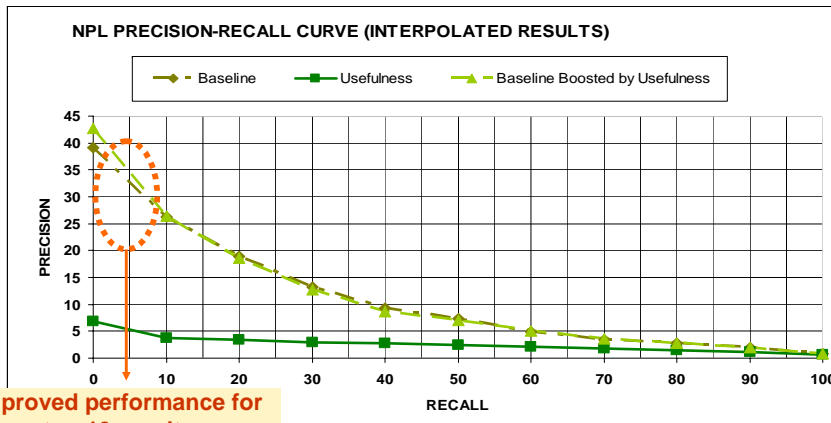
## Experimental Setup

- Dataset: NPL collection
  - 30 experimental queries
  - 10,737 query-relevant documents
  - Every NPL document approximates the snippet size (~23 terms)
  - NPL queries vary in size between 2 and 9 words
- Getting Started
  - Semi-automatic annotation of queries with an appropriate WordNet sense
  - Semantic annotation of all document content terms
  - Computation of semantic similarity values between query and document terms

## Experimental Setup (2)

- Merge NPL documents into a single text (virtual document)
- Issue queries and select snippets
- Comparison of snippets selected by the TF/IDF statistical model to the snippets selected by our semantically-driven model and the snippets selected by their combination
- Evaluation metric: interpolated 11-point Precision-Recall curves

## Experimental results



Improved performance for top 10 results

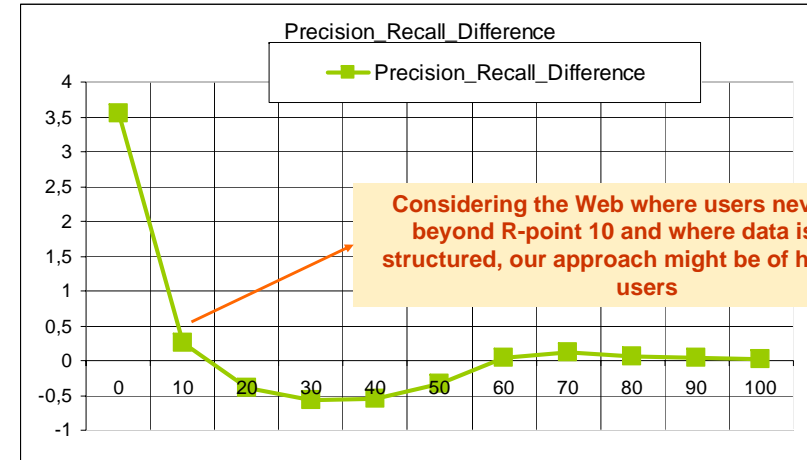
The combination of statistical and semantic criteria for snippet selection yields a 3.5% improvement

9/8/2007

SIGIR-2007-FOCUS

17

## Experimental Results (2)



Considering the Web where users never reach beyond R-point 10 and where data is semi-structured, our approach might be of help to the users

Improvement is non-negligible considering that NPL is a well-structured, balanced and small data collection

## Human Survey

Recruited 15 users and used the NPL dataset to select snippets based on:

- Baseline Alicante algorithm
- Usefulness metric
- Semantic Coherence metric
- Expressiveness metric

Which of the displayed snippets do you think will direct you to a document that can successfully answer the query intention?

Queries were disambiguated and their selected senses were displayed

USER	Baseline	Query Usefulness	Semantic Coherence	Text-Expressiveness
1	5	12	9	8
2	9	17	8	5
3	6	7	7	10
4	8	17	9	10
5	8	13	7	5
6	11	15	5	6
7	3	15	7	6
8	14	14	4	3
9	9	9	10	7
10	11	15	5	6
11	4	11	9	6
12	7	15	11	8
13	5	10	11	9
14	9	18	6	8
15	6	12	7	5

Users prefer Query-Relevant Snippets

Semantically-derived snippets are valued higher

Semantics-driven passage retrieval can assist the users focus on retrieved results

## Conclusions

- **New approach** for query-centric snippet selection
- Evaluation models for measuring snippet **coherence** and **expressiveness**
- Extensive **experimentation** will help us define the contribution of every metric in the snippet selection process
- A novel technique towards **personalized passage retrieval algorithms**

9/8/2007

SIGIR-2007-FOCUS

20

---

Thank You😊

**QUESTIONS?**