Kyoto Semantic Search and User Evaluation

Feikje Hielkema
Irion Technologies
Piek Vossen
VU University Amsterdam

Contents

- Introduction
- From text-based to conceptual search: the three Kyoto search systems
- Comparing search methods through evaluation
- Discussion & Conclusion

Introduction

- Aims:
  - Develop a search system that provides access to valuable information across languages, cultures and media, through deep semantic analysis of textual information.
  - Evaluate the system in terms of usability and usefulness in comparison to simpler and more familiar text-based search systems.

From Text-based to Conceptual Search

- Kyoto has developed three search systems:
  - The Baseline: its text-based results are presented as a list with snippets and a relevance score.
  - Semantic Search, which finds results with Baseline, but extracts approximation of facts from the search results and provides different views (e.g. map and table).
  - Conceptual Search, which finds results from indexed facts through matching concepts, and presents them as facts with different views.

The Baseline System

- Based on the TwentyOne Search system developed by Irion Technologies.
- Phrase matching based on:
  - The proportion of query words that are included in the phrase;
  - The degree to which the query words match the phrase words;
  - Using synonyms, fuzzy matching, compound and multiword inclusion.

The Baseline System

- Results are presented in a list, with snippet and relevance score.
- Supports cross-lingual search for English, Dutch, Spanish, Basque, Italian, German & Japanese.
- Demonstration.

Semantic Search System

- Identical phrase matching (using the same TwentyOne Search software);
- The system uses the KAF-files to extract properties, quantities, locations and dates from the context of these phrases;
  - Locations & dates are marked in the KAF during NER-extraction;
  - Properties, quantities and location types (e.g. moor, coast) are extracted using word lists.

Semantic Search System

- These 'facts' are presented in a Simile Exhibit (http://www.simile-widgets.org/exhibit/)
  - Includes three different views: table, tiles & Google map;
  - Results can be filtered and sorted by their various *facets* (i.e. property, location, date).
- Demonstration

Conceptual Search

- Analyses the textual query to a set of concepts;
- Searches in the collection of facts extracted by Kybots (see 'Mining events and facts in Kyoto', German Rigau and Aitor Soroa, tomorrow);
- Extracts all facts with these concepts;
- Orders them by the strength and number of matches;
- Displays the results in a Simile Exhibit.

Example of indexed fact:

<event eid="e40" lemma="unpolluted" pos="G" target="t2261" synset="eng-30-01907711-a" rank="1.0">

<role rid="r44" event="e40" target="t2255" lemma="water" pos="N" rtype="patient" synset="eng-30-14845743-n" rank="0.244333"/>

<role rid="r45" event="e40" target="t2260" lemma="largely" pos="A" rtype="state-of" synset="eng-30-00006105-r" rank="0.516245"/>

<place countryCode="US" countryName="United States" name="Atlantic" fname="populated place" latitude="41.4036007" longitude="-95.0138776">

</place>

<dateInfo dateIso="1999" lemma="1999">

</dateInfo>

</event>

Analysing the Search Term

- Using a term database, the system identifies a set of concepts by lemma and pos-tags;
  - habitat of king penguins → habitat-n + king_penguin-n.
- These are disambiguated and expanded by the Word Sense Disambiguation by Evocation service to a set of synset-ids;
  - Each synset has a confidence score.
- These synsets are expanded, using Wordnet, with their hypernyms.
  - The further removed the hypernym from the synset, the lower its confidence score.

Indexing the Kybot Facts

- Facts are indexed by:
  - Lemma;
  - Synset ID;
  - Synset ID of hypernyms.
- Facts are indexed with:
  - Lemma's & synset IDs, with confidence value;
  - Reference to page in original document, and context sentence;
  - Locations & dates, for presentation on map.

Retrieving Kybot Facts

- Retrieve all facts which:
  - Have a synset which matches a synset or hypernym from the analysed query;
  - Have a hypernym which matches a synset from the analysed query.
  - Have a lemma which matches a query lemma.
- Order them by relevance score:
  - The sum of the score of all matches between query & fact;
  - The score of each match is the product of its synset's confidence values.

Conceptual Search

- The Conceptual Search System thus matches concepts, rather than phrases, and presents facts, rather than snippets.

- Demonstration

Comparing Search Methods through Evaluation

- In the course of their work, users search for answers to complex questions.
  - E.g. What is the impact of declining bee populations on agricultural productivity?
- Which tool supports this task best - Text-based or Concept-based?
- We have compared the three Kyoto-tools in a task-based experiment.
  - Each tool searches in the same database;
  - Baseline and Semantic Search search identically;
  - Semantic and Conceptual Search present identically.

Evaluation - Methodology

- 20 subjects:
  - 4 environmental professionals at ECNC, 6 students of environmental sciences and 10 students of various Arts disciplines at the VU.
- Answer 6 high-level questions with each tool.
  - Open questions, answers must be phrased in text;
  - Answers are lists, and must be found in different documents to be complete.
- Feedback was gathered using the System Usability Scale (Brooke, J. ,1996), and a comparative questionnaire at the end of the experiment.

SUS Questionare

1. I think that I would like to use this system frequently
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system

Evaluation - Methodology

- We measured:
  - Time needed per question;
  - Number of searches per tool (=6 questions);
  - Number of documents viewed per tool;
  - Number of correct answers:
    - Strict form: incomplete or partially correct = incorrect;
    - Lax form: incomplete or partially correct = correct.

Evaluation - Methodology

- Each subject used each tool, and answered three different sets of questions;
  - The order and combination of tools and question sets were varied to avoid training effects;
  - Each question must be answered in 10 min.
- Before receiving a question set, each subject worked through a one-page introduction to the next tool.
- The experiment lasted between 3 and 4 hours.

Evaluation - Hypothesis

- Null hypothesis: subjects will find equally accurate with each tool, using the same number of search terms, viewing the same number of documents in the same length of time.

- Research hypothesis: Subjects will be more complete in the answers found using the Conceptual Search system than in the other two, using less searches and viewing less documents.

	Benchmark	Text-based facts	Conceptual Search	ANOVA Bonferroni post-hoc test (1&2; 1&3; 2&3)
Time per question	μ = 405, σ = 125	μ = 450, σ = 65	M = 482, σ = 70	.070; .033; .148
Correct answers	μ = 2.30, σ = 1.17	μ = 1.80, σ = 1.32	μ = 1.50, σ = 1.28	No differences between groups
Partially correct answers	μ = 4.95, σ = .83	μ = 4.40, σ = 1.43	μ = 4.15, σ = 1.35	No differences between groups
Searches	μ = 31.1, σ = 13.11	μ = 24.6, σ = 8.31	μ = 21.4,	.092; .173; 1.00
Documents viewed	μ = 21.5, σ = 8.28	μ = 23.4, σ = 6.53	μ = 21.9, σ = 7.02	No differences between groups
SUS	μ = 71.1, σ = 15.27	μ = 58.2, σ = 19.17	μ = 52.0, σ = 20.82	.063 ; **.006**; .958

Evaluation - Results

- Significant difference in SUS-score between Baseline and Conceptual search, in favour of the Baseline.
- No significant differences in correctness or completeness of the answers.
- No significant differences in time, search requests and viewed documents.
- Conclusion: subjects were approx. equally effective with each tool, but preferred the Baseline. Why?

Evaluation - Feedback

- 10 Users liked the Baseline:
  - user friendly
  - simple design
  - more like the conventional 'Google' idea
- And were baffled by Conceptual Search:
  - Could not find word matches (the thing you normally search with/for);
  - I was very confused by the columns
  - I didn't understand the terms 'patient' or 'simple cause',
  - Lots of technical jargon in table.

Evaluation - Feedback

- 6 users liked Conceptual Search:
  - I liked that the system could recognize causal relationships
  - I liked this system best as it allowed me to adapt my search using the facets
  - It was possible to enter an entire question, this method mostly worked and provided more specific results
- And disliked the Baseline:
  - You had to be very specific with the search words
  - The findings were difficult to sort out

Evaluation - Discussion

- The more powerful functionality of Conceptual Search decreases its usability and learnability.
- Users who wish to search immediately, and not spend time learning to use the system, will prefer the more 'Google-like' Baseline.
- However, Conceptual Search is liked by more 'adventurous' users, who will investigate the extra functionality if they believe it will help them to search more effectively in the end.

Evaluation - Discussion

- What can we do to make Conceptual Search less daunting to novice users?
  - Show why each search result is found; e.g. by highlighting which concepts have matched the search term, and/or displaying the concepts to which the search term was interpreted. Loss of confidence in the search results is lethal to any search system.
  - The 'cause' and 'patient' tags are often not understood by users, or do not match their expectations due to errors in the facts.

Evaluation - Discussion

- What can we do to make Conceptual Search less daunting to novice users?
  - We need to present the facts in a way that users understand. Context, locations and dates were clear; but actor/cause and patient/result were found confusing by many.
  - We need greater accuracy in our facts; when users are struggling to understand the meaning of 'event' or 'patient', events like 'crab' or 'shark' will mislead them, which will hamper their understanding of the other facts.

Conclusion

- Conceptual Search must be improved in terms of usability;
- It must be improved in terms of accuracy:
  - We need greater precision and recall in both the kybot facts and the query disambiguation.
- Although it baffled many users, their answers were neither more nor less accurate or complete.
  - If we make it clearer to users why they see particular search results, and increase the confidence in these results, the greater usability may increase the effectiveness as well.