Lab 5: Information Access

TDT4275: Natural Language Interfaces

Jonas Myrlund

April 23, 2013

1 Written Assignments

A Using either web searches or a large corpus, devise a query using a polysemous word.

I want to find out more about large cranes – the bird. I start out with the following query: "large crane".

See figure 1 for the initial results using Google[1].

B Then, devise two or three alternate queries which help to disambiguate the polysemous word and repeat the search.

The query "large crane" could be less ambiguous in the following ways:

1. "large crane bird"
2. "large crane feathers"

As figure 2 exemplifies, these queries result in only documents relating to the *bird sense* of the word *crane*.

C For each query, classify each of the top 10 search results as correct or incorrect.

Again, I will be using Google for the queries.

C.1 Query: "large crane"

The query yields only one correct document: the second search result. The others documents relate to the wrong sense of the word. For details, see figure 1.

[1] https://google.com/search?q=large+crane

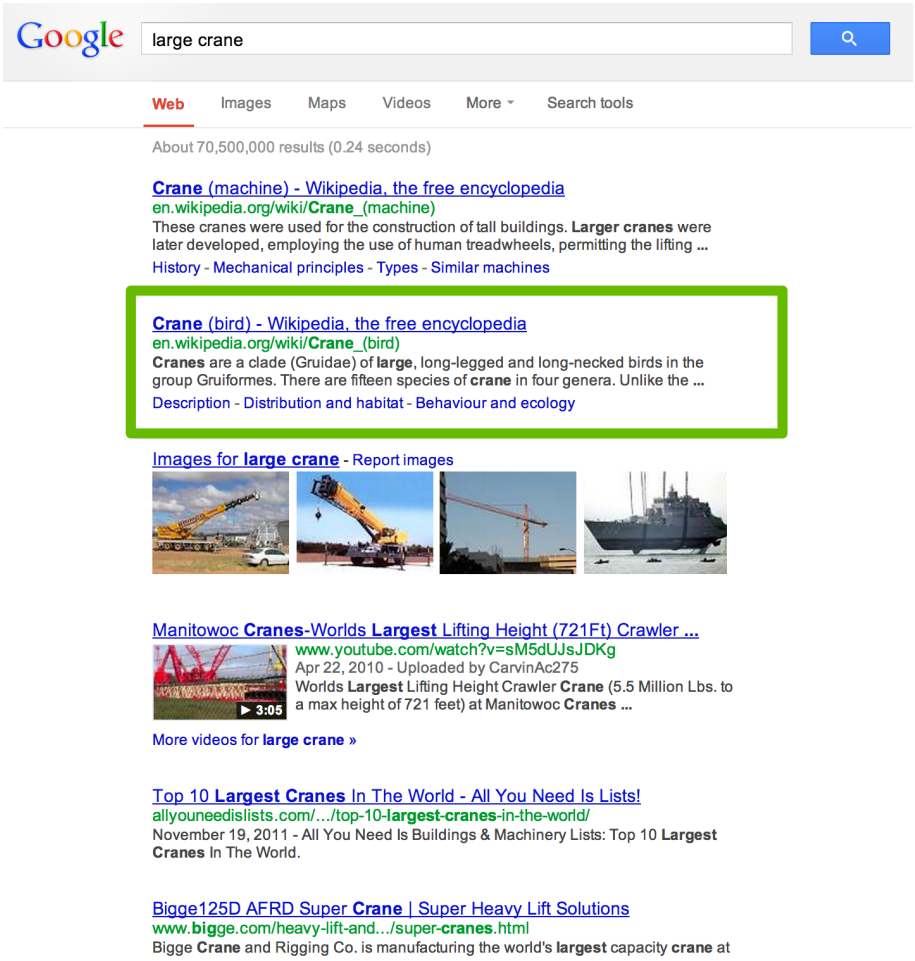

Figure 1: Google yields only one relevant search result when searching for "large crane".

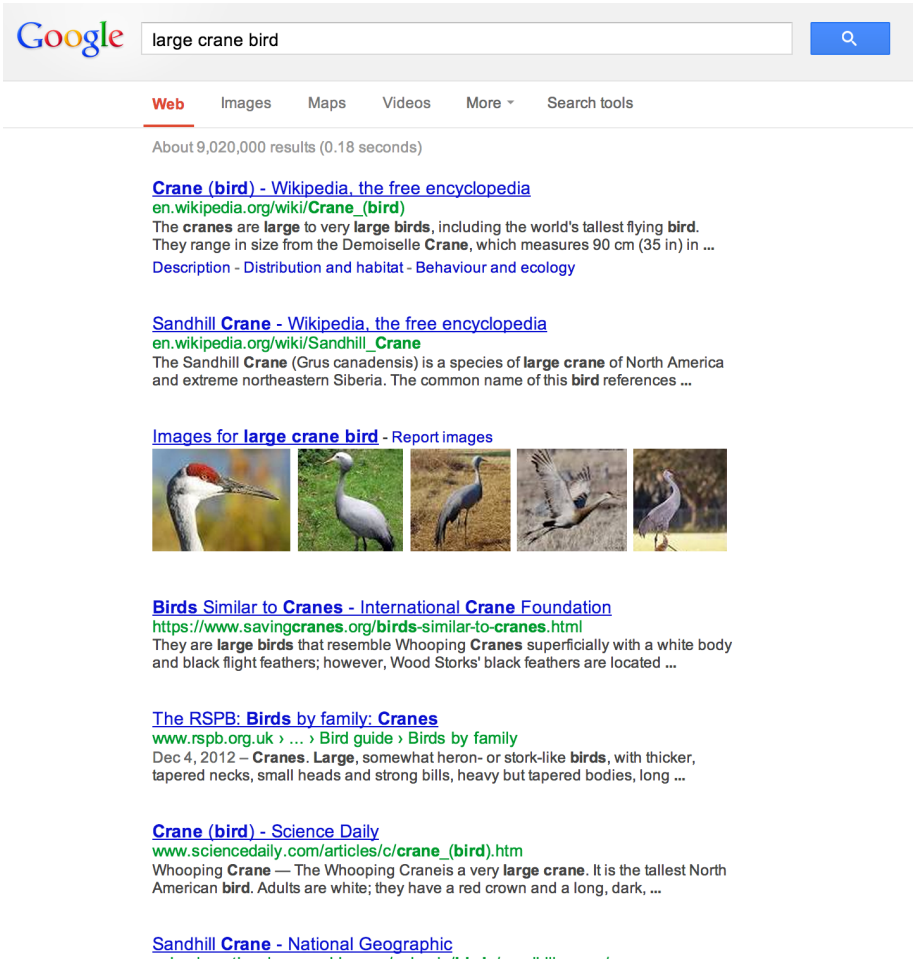

Figure 2: When adding the term "bird" to the query, all the top results are relevant.

C.2 Queries: "large crane bird" and "large crane feathers"

For both these queries, all the first 10 search results relate to the right sense of the polysemous word in Google.

Yahoo! and Bing, however, both give three documents relating only to feathers, which is not what is wanted.

D Create an interpolated precision-recall curve comparing the accuracy of each of the queries.

The recall values for the first query, "large crane", stagnates in that it gives only one correct document. Thus, the scales of the precision-recall curves do not overlap to a degree that makes it viable to plot them together.

However, the curves quite clearly show the improvement adding a disambiguating term yields, when the precision axis is normalized to one. See figure 3, 4 and 5.

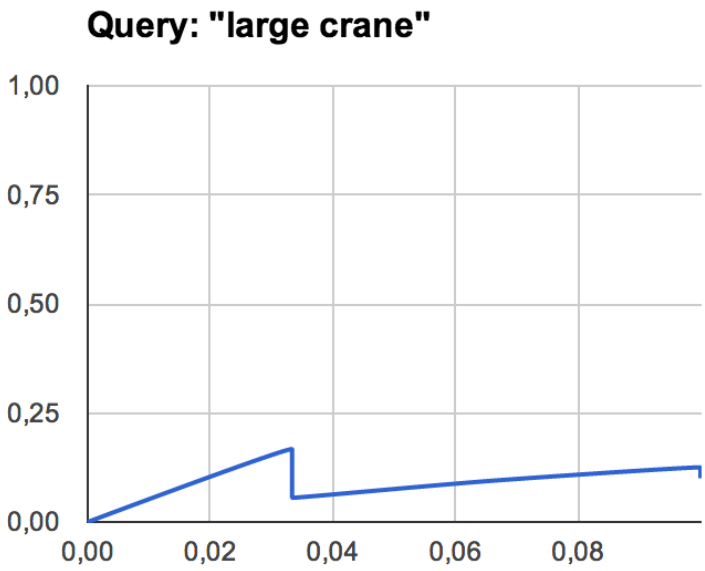

Figure 3: Precision-recall curve for "large crane".

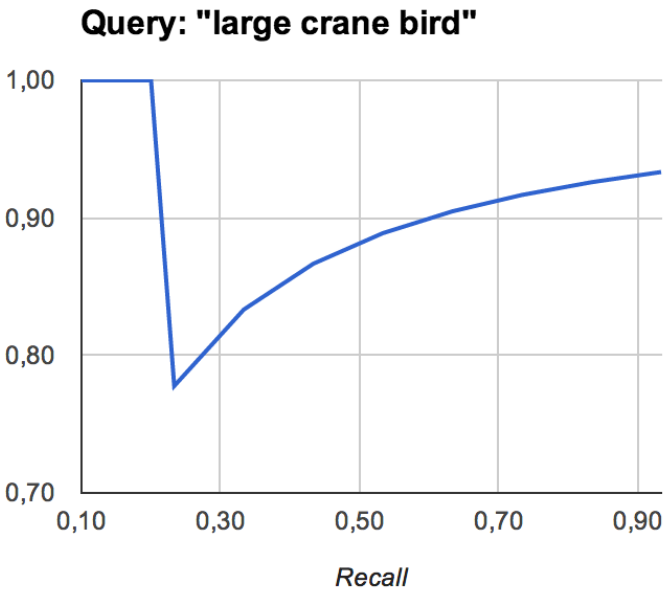

Figure 4: Precision-recall curve for "large crane bird".

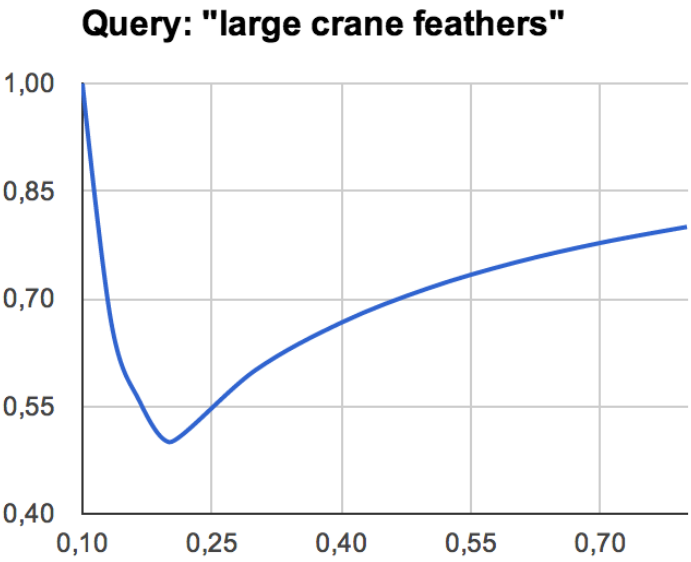

Figure 5: Precision-recall curve for "large crane feathers".

2 Information Retrieval

I went for a corpus traversed with Lynx from http://www.sciencenews.org/.

A The effects of the various steps described in the assignment.

Turn the corpus into text form. This really has no significant effect, as the corpus garnered from Lynx is already stripped of all HTML tags.

Turn the file into a list of words. Removes special characters, lowercases, and generally simplifies words into tokens, like this:

Before	Animals' cognitive shortcomings are as revealing as their genius.
After	animals cognitive shortcomings are as revealing as their genius

Build word frequency tables. This results in a series of *term frequency* files, of the following format:

```
0.00342987804878049     animals
0.00114503816793893     cognitive
0.00114503816793893     shortcomings
```

The number to the left denotes the term frequency of the word to the right.

Create your own stop list. This step is simply a matter of looking manually at some of the generated files and seeing which words are present in every document. These are manually extracted into a stop list.

The beginning of my stop list looks like the following:

```
a
all
the
of
and
in
for
on
or
```

Determine the terms' descriminative values. This step calculates the IDF – the inverse document frequency – of each word found in the tokenized document collection. The IDF routine simply outputs each word together with the inverse of the term frequency over the entire corpus.

Experiment with a simple search engine. Searching for "physical society" should probably lead to a document summarizing a recent meeting in the Physical Society.

Running `echo physical society | ../bin/search.perl -i idflist -e tfs` results in the following (please note that the numbers to the left are a result of some heavy playing around with weighting TF and IDF terms):

```
2430.451300474519204 lnk00000036.tfs
0.451300474519204 lnk00000037.tfs
0.204453275655133 lnk00000030.tfs
```

As expected, the raw document `lnk00000036.txt` is the document we were looking for.

Using a stoplist has no significant effect with the configuration used in the example above.

3 Text Summarisation

A

The SweSum summariser[2] is first and foremost a single DOC summariser. Furthermore, it accepts keywords to aid in the summarisation, in this way handling a sort of QUERY input as well.

The summariser is an EXTRACT summariser – it selects what it deems the most relevant sentences, and concatenates them together.

B

The first article I've chosen to summarise is one about seafowl being killed by seismic shooting under water[3]. The summaries are attached in appendix A-1.

The second article is a short sports article[4]. The summaries are attached in appendix A-2.

[2] http://swesum.nada.kth.se/index-eng-adv.html
[3] http://www.nrk.no/nyheter/distrikt/troms_og_finnmark/1.10996526
[4] http://www.nrk.no/sport/fotball/tyson-stotter-suarez-1.10997215

C

I would rank the summaries in the following way:

1. 25% version of article A-2.
2. 50% version of article A-1.
3. 50% version of article A-2.
4. 25% version of article A-1.

My favorite summary is extremely short. However, it highlights the important parts of the article, while leaving out a lot of the quite uninteresting parts.

The real story – the message – is in the first couple of paragraphs, whereas the latter half of the article raves about how the two people involved heard about each other and started following each other on Twitter – very peripheral material. This last bit is all left out in the summary, and would probably never be missed by anyone.

The worst summary in my opinion fails to acknowledge the different opinions within the original article. The article is split in two, the first half presenting one view, the other half presenting an opposing view, and it does not conclude one way or another. The summary, however, presents only the first view, and fails to even mention the controversy highlighted by the latter half of the article.

I can think of two ways of making sure the omitted sentences are included: to *identify the headlines* and include at least one central sentence from each of these headlined sections, and *classifying the names of people involved*, ensuring that one quote from each interviewee is included in the summary.

A Text summaries

A-1 Article 1 – "Mener lundefuglene kan ha dødd på grunn av seismikkskyting"

The original article can be found on NRK.no.

The *25% summary*:

> I forrige uke fant barnehageunger på tur i Botnfjæra i Dyfjord i Finnmark 11 døde lundefugler. Fisker og sekretær i Andøy Fiskarlag, Bjørnar Nicolaysen, mener konsekvensene av seismikkskytingene i Vesterålen, Lofoten og Senja er mye verre enn antatt.

De døde lundefuglene som nylig ble funnet i fjæra i Dyfjord i Lebesby kommune i Finnmark kan skyldes trykkbølgene fra seismikkskyting, frykter Nicolaysen.

– Lundefugl som annen sjøfugl er avhengig av maten som finnes under havflaten.

Seismikkskyting er oljeindustriens viktigste redskap for å kartlegge mulige forekomster av olje og gass mange tusen meter under jordoverflaten, og er derfor definert som petroleumsvirksomhet i petroleumsloven. Det er blitt undersøkt konsekvensene av effektene av seismikk på havet.

The *50% summary*:

I forrige uke fant barnehageunger på tur i Botnfjæra i Dyfjord i Finnmark 11 døde lundefugler. Fisker og sekretær i Andøy Fiskarlag, Bjørnar Nicolaysen, mener konsekvensene av seismikkskytingene i Vesterålen, Lofoten og Senja er mye verre enn antatt.

De døde lundefuglene som nylig ble funnet i fjæra i Dyfjord i Lebesby kommune i Finnmark kan skyldes trykkbølgene fra seismikkskyting, frykter Nicolaysen.

– Lundefugl som annen sjøfugl er avhengig av maten som finnes under havflaten. Dukker den ned der det skytes seismikk, så er den ferdig.

Fisker Bjørnar Nicolaisen fra Andøy mener lundefuglene kan ha dødd på grunn av seismikkskyting.

Seismikkskyting er oljeindustriens viktigste redskap for å kartlegge mulige forekomster av olje og gass mange tusen meter under jordoverflaten, og er derfor definert som petroleumsvirksomhet i petroleumsloven.

– Det er ingen av oss som kjenner til episoder hvor seismikk har drept sjøfugl. Tvert om antar vi at sjøfugl svømmer under båten når den nærmere seg, og blir ikke påvirket i det hele tatt.

– Jeg tviler på at lundefuglepisoden har noe med seismikk å gjøre.

– Kan omfattende seismikkskyting påvirke fuglelivet? Det er blitt undersøkt konsekvensene av effektene av seismikk på havet. Det ble også gjort undersøkelser på sjøpattedyr.

De har også deltat i overvåkings- og kartleggingsprogram for norske sjøfugler (Seapop), hvor seismikk og sjøfugl aldri tidligere har vært et tema.

A-2 Article 2 – "Bite-Tyson tar bite-Suarez i forsvar"

The original article can be found on NRK.no.

The *25% summary*:

Bite-Tyson tar bite-Suarez i forsvar – Han bet noen, sånt skjer, sier Mike Tyson.

Etter Luis Suarez' siste skandale, søndagens biting av Chelsea-spiller Branislav Ivanovic, får han støtte fra noen som virkelig kan sette seg inn i situasjonen hans - bokselegenden Mike Tyson.

Mange husker nok da Tyson satte tennene i Evander Holyfields øre under en boksekamp i 1997.

The *50% summary*:

Bite-Tyson tar bite-Suarez i forsvar – Han bet noen, sånt skjer, sier Mike Tyson.

Mike Tyson uttalte seg om Luis Suarez i et radioshow tirsdag.

Etter Luis Suarez' siste skandale, søndagens biting av Chelsea-spiller Branislav Ivanovic, får han støtte fra noen som virkelig kan sette seg inn i situasjonen hans - bokselegenden Mike Tyson. De er og var begge i verdenstoppen i sine idretter, og begge har brukt biting i forsøk på å stanse sine motstandere.

Mange husker nok da Tyson satte tennene i Evander Holyfields øre under en boksekamp i 1997.

Følger Suarez på Twitter

Etter biteepisoden søndag var det flere som merket seg at plutselig hadde bokselegenden begynt å "følge" Suarez på nettstedet Twitter.

Tyson bekrefter at det var bitingen som gjorde ham oppmerksom på Liverpool-spilleren. Han kan vente seg flere kampers karantene.