

Will Powerset be here in 12 months time?

Marek Foss, 06/04/2008

Powerset[1] declared it aims at Google[2] - the company that revolutionised web search and became the icon of Internet. Today, we do not ask questions, we google answers. Still, lots of improvements are possible. For example, searching for academic papers on a particular topic in Google is a task of correctly transforming the question into keywords - even a reordering of the query terms gives fairly different results, and often the most relevant documents appear below the top ones. Powerset target is to transform search from a keyword-based into a natural language question answering. The idea is to enable users ask questions, like they normally would ask other people, and answer them with high correctness[3]. Thus, not only natural language processing (NLP) of queries, but also NLP of web resources is involved (although now, Powerset just uses Freebase[4] and WordNet[5]). In other words, instead of analysing documents for terms and their rank, the system will analyse the linguistic structure of it, what classes the terms fall into (like *company*, *person* etc.) and combine this knowledge into sets (for example, personal bio files).

Powerset technology basis on a licensed work from Xerox PARC laboratories[6], which spent 30 years on developing and enhancing NLP systems. Powerset idea of natural language search consists of four elements[7]:

- interpreting the web
- indexing
- interpreting the query
- searching and matching

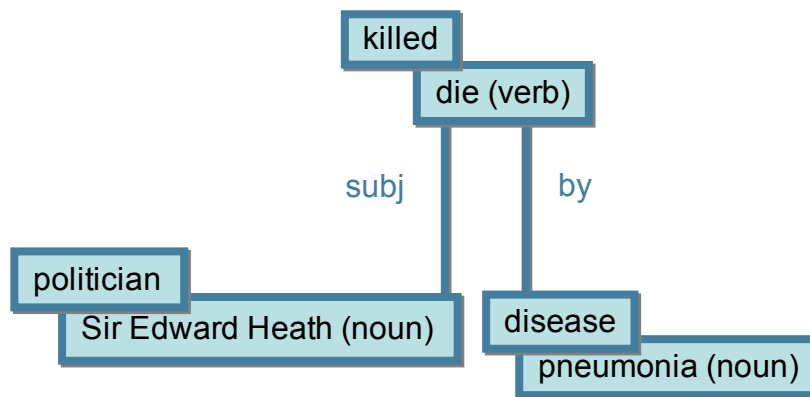


Fig. 1. Analysing a document sentence[7].

In interpreting the web, Powerset analyses documents by processing each sentence on several levels. Firstly, sentence entities as parts of speech are extracted, assigning nouns, verbs, adjectives etc. to words. Also, semantic relationships are analysed, giving connections i.e. between words, like a noun is a subject of a verb etc. For example, in "*Sir Edward Heath died from pneumonia*"[Fig. 1], *Sir Edward Heath* is classified as a noun phrase, *pneumonia* is also a noun, while *died* is a verb. Preposition *from* is encoded as a relation 'verb by noun', '*died* by pneumonia'.

Sir Edward Heath is a subject of that verb. In fact, Powerset uses more complex parser, with results similar to the box output of CCG parser[Fig. 2].

Secondly, the extracted information is further analysed and generalised - entities are assigned to similar abstract classes, gaining relationship. For example, using Freebase knowledge, *Sir Edward Heath* is classified as a *person*, and by increasing the specialisation - a *politician*. With WordNet data, *pneumonia* is classified as a *disease*, while the verb *died* is associated with synonyms like *killed*. At the end, indexing collects multiple facts for each sentence, that where generalised through previous analysis.

```

[PRED 'parse<[22:we], [70:Web.]>'
SUBJ [
  [PRED 'we'
  NTYPE [NSYN pronoun]
  22[CASE nom, HUMAN +, NUM pl, PERS 1, PRON-TYPE pers]
  [PRED 'Web'
  NTYPE [NSYN proper]
  OBJ [
    SPEC [
      [DET [PRED 'the'
      DET-TYPE def]
      70[CASE obl, NUM sg, PERS 3]
    ]
  ]
  TNS-ASP [MOOD indicative, PERF -, PROG -, TENSE pres]
  42[CLAUSE-TYPE decl, PASSIVE -, VTYPE main]
]

```

Fig. 2. Powerset parser output[7].

Using such approach, the Powerset system is able to answer questions like "*What killed Sir Edward Heath?*" - interpreting the query is a similar process to interpreting the web. It can be viewed as growing trees from each query term into leaves of generalisation. Similar trees grow from terms of documents sentences, and eventually their leaves will meet with some leaves from the query, giving the desired answers[Fig. 3]. It is a very different approach from a straightforward keyword matching, which is the core technology of current search engines.

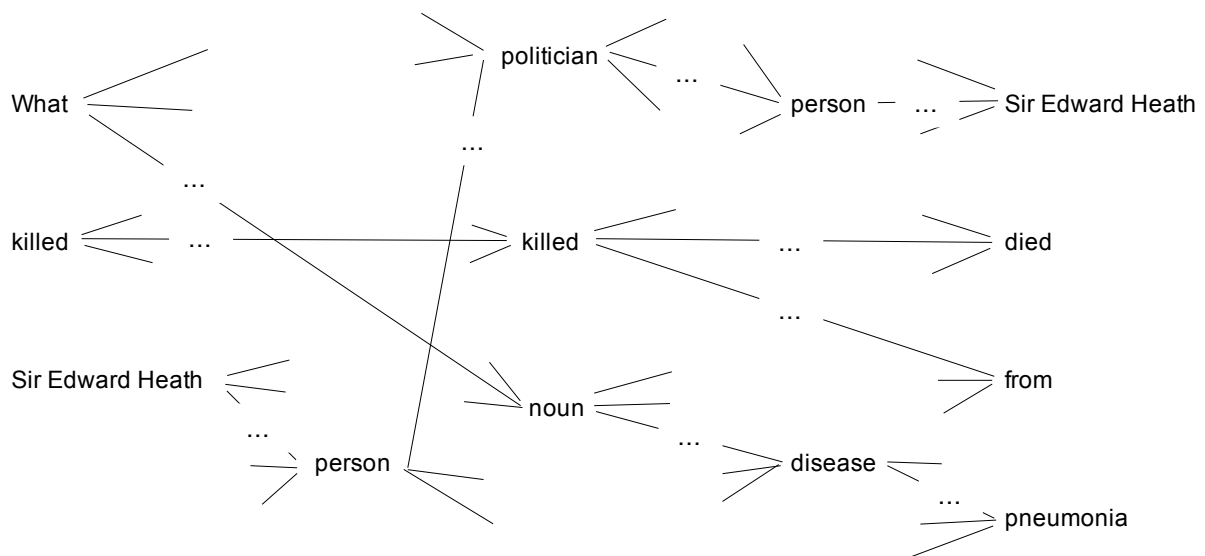


Fig. 3. Searching and matching of query and resources.

Therefore, the user experience differs a lot. Firstly, one has to formulate a question, instead of just typing word relevant to the search subject. Secondly, the margin of modifying the question, in case of not satisfying results, is very limited (as opposed to simple query terms reordering mentioned earlier). In fact, my tests on the current demo available on Powerset closed beta, the Wikipedia Search Sneak Peek[8], shows great difficulties in formulating questions. The constant use of Google created a habit of formulating problems as single words, no longer that 3-5 term queries. Another problem is input overhead, because Powerset requires fairly grammatical question, with wh- terms, verbs etc. Thus, we need to type "*What is the best perl framework for web applications*". It is true that simplification to "*what best perl framework web applications*" does not change the results, but then what is the point of NLP, and why not to just use "*best perl framework web applications*" in Google? Results are better, pointing straightforward to framework websites (Powerset may have worse results due to analysing only Wikipedia content[9], still each of the frameworks in the top Google results has a Wikipedia entry).

On the other hand, some questions show marginal advantage of Powerset, for example "*can noun be a subject of a verb?*" In this case, the natural language search resulted with a citation "*In playboy, for example, the noun is the subject of the verb (the boy plays)*". Google does not give such a direct answer when queried "*noun is subject of a verb*", however situation changes when we put the query in quotation marks - we get a link to *Tricky Nouns* with citation "*if a noun is the subject of a verb, which verb form to use?*". Trying questions in Google also is interesting - "*can noun be subject of a verb*" results in "*A noun can function in a sentence as a subject*".

The key problem with search system proposed by Powerset is the lack of flexibility. It is not user-friendly to force question-based querying and return Internal Server Errors on keyword-based queries. Drawing a line between natural language and keywords is artificial and made on the basis of idealism rather than usability. Moreover, it is not enough to base on Freebase and WordNet knowledge. The current approach makes the system more of a world knowledge search, an encyclopaedia, rather than a web search engine. Web crawling and parsing is essential. Will Powerset be here in 12 months and appear as a revolution of 2009, or will it vanish into the internet cache? Its public beta is launching in upcoming weeks, but in the current state it is more of a scientific experiment on NLP than a useful tool for masses. Flexibility of querying, fine-tuning of output and improving user interface seems to be the most important challenges Powerset has to deal with, to stand in front of Google as an opponent rather than a curiosity.

References and Links:

[1] Powerset Inc. (<http://www.powerset.com>)

[2] Google Inc. (<http://www.google.com>)

[3] "Search startup Powerset says it can knock off Google", Richard Koman, Silicon Valley Watcher, (http://www.siliconvalleywatcher.com/mt/archives/2006/10/search_startup.php)

[4] Freebase, (<http://www.freebase.com>)

[5] WordNet, (<http://wordnet.princeton.edu>)

[6] "Powering up at Powerset", Verne Kopytoff, The Technology Chronicles, SFGate.com, (http://www.sfgate.com/cgi-bin/blogs/sfgate/detail?blogid=19&entry_id=18129)

[7] "Powerset - Natural Language and the Semantic Web", Barney Pell, The 6th International Semantic Web Conference

and the 2nd Asian Semantic Web Conference 2007, (http://videolectures.net/iswc07_pell_nlpsw/)

[8] Wiki Search Sneak Peek, Powerset Inc. (<http://labs.powerset.com/demos/Wiki+Search+Sneak+Peek>)

[9] "Let's Test Powerset", Ionut Alex Chitu, Google Operating Systems Unofficial Blog, (<http://googlesystem.blogspot.com/2007/12/lets-test-powerset.html>)