How language technology is making Danish search engines smarter

Language technology will revolutionise search engines by teaching them language to deliver not just references to documents but full, valid answers in our natural idiom.

Maria Bernbom

Why today's search engines are dumb

"21 documents match your search…" Fantastic! Another couple of clicks and you find the information you were looking for. Five years ago, the same piece of information would have cost you three phone book references, two phone calls and half an hour of your precious time. In 2002 you can find what you seek in 48 seconds. But only if you're lucky. Current engines annoy us because they don't answer our questions – only point us to a plethora of documents.

Increasingly, we need answers to specific, apparently simple questions such as

- "How do I get rid of bulky household refuse?"
- "When will the film *The Two Towers* be released in the UK?"
- "My company wants to install water coolers. Who is the cheapest supplier?"

Even though this information is available somewhere on the Internet, even the best search engines are not "intelligent" enough to find it, process it and present it to us in the most appropriate form –"*The Two Towers* will be released on Wednesday 18 December 2002". This is because search engines do not *understand* what we are asking for – they only *track* the words we use.

Which is where language technology gets into the picture.

The internet must learn language

Language technology is being developed that will make the Internet understand us in our own natural language. It will teach a search engine that a question like *"Which car dealer in East London sells Jaguars?"* means that we want to know where in a certain area we can buy a Jaguar car, but we aren't interested in car dealers in the North West of England nor in the jaguars of the animal kingdom.

In other words, search engines must learn the various possible meanings of words and sentences. Which is why researchers at universities and in industry all over the world are already working hard on the development of more intelligent search engines. In Denmark this research is being carried out in the following interesting projects:

- MOSES – Towards a more intelligent Internet **[**LINK]
- OntoQuery – Because thiamine is also a vitamin **[**LINK**]**
- Ankiro – When businesses train search engines and chat robots **[**LINK]

MOSES – Towards a more intelligent Internet

"The web of today is a collection of texts that people can read and use. We see the web of the future as "a semantic web", which can of course still be read, but which can also find the particular answers we seek automatically – and above all independently of the actual natural language we use for our search," says Patrizia Paggio, senior researcher at The Centre for Language Technology (CST) in Copenhagen. CST is a partner in the European MOSES project, together with the universities of Rome and Copenhagen and a number of software companies in Italy, France and Holland. This is a project primarily aimed at investigating two issues:

- How can we "label" the texts that we publish on the internet to make it easier for the search engines to understand what they contain? **[**LINK to Keywords to summarise the web page content**]**
- How can we make search engines search the Internet more intelligently? **[**LINK to Search engines must understand languages]

Keywords to summarise the web page content

Just as libraries label every book with a range of keywords that indicate the subject matter of the book, texts on the web will also be labelled or 'marked up' with keywords to help evaluate their relevance to answering a specific query.

This obviously means that web users will have to mark up their web pages when they publish them on the Internet. Patrizia Paggio doesn't think this should cause much trouble: "Today we already have to define formatting information such as font style and size when uploading a text to the Internet; in future, we will simply add a few keywords indicating the subject matter of our text".

What MOSES is trying to develop is a way of doing this 'semantic' mark up automatically. Paggio's vision is that the search engines will probably become so intelligent that they can themselves suggest a relevant semantic mark-up for our texts, and we will simply validate the system's suggestions.

Search engines must understand language

To ensure that automatic mark-up becomes a reality, a search engine will have to do more than recognise a range of keywords in a text. It will also need to know that homonyms - words spelt in the same way such as 'bank' (finance) and 'bank' (river) - have different meanings, and that some words with very different spellings can have the same or a closely related meaning (e.g. the verbs 'open' and 'inaugurate').

"In our project, we will mark up parts of the web pages for the universities of Rome and Copenhagen," says Patrizia Paggio. "A student might for example ask the search engine: 'What courses are offered in medieval art in the spring term 2002?' The search engine may not be able to find any 'courses' on medieval art as such in the database, but it will probably find a "lecture", a "seminar", a "workshop" or a "study group" on the subject. These words are near synonym to 'course' in this context, and might therefore interest the searcher."

To achieve this, someone has to tell the search engine that these words have similar meanings. Which means that we have to build the relevant ontologies – or conceptual systems which relate words and concepts to each other, so that the search engines know that a 'course' and a 'seminar' are both related to teaching in this context.

OntoQuery – Because thiamine is also a vitamin

Such a conceptual system is exactly what the researchers on the Danish OntoQuery project are trying to develop. They have developed a prototype for an ontology-based search engine [LINK to http://www.ontoquery.dk], which can be used when searching for nutrition-related subjects in *The Danish Encyclopaedia*.

"Our search engine must be able to recognise the same content in different disguises," explains Bolette Sandford Pedersen, a senior researcher at CST, one of the partners in this project. For example, the engine must know that there is a certain relation between the following concepts:

- Lack of vitamin
- Vitamin deficiency
- Lack of thiamine
- Malnutrition

In other words, the researchers have to teach the engine how meanings are related through words. Current search engines use the techniques of pattern recognition (i.e. letters and numbers in certain combinations) and statistics (web pages linked to and visited the most, and where and how often our search words are mentioned in the document) to find documents relevant to a query. But more advanced search engines will have to know how concepts are related to each other, and whether these concepts are valid synonyms or are super- or subordinate to each other.

"It is important, especially for laymen, that the engines are able to recognise concepts however they are presented," Bolette Sandford Pedersen emphasises. Experts are often able to ask relatively precise questions, whereas non-specialists typically use more general search terms. For example, they may not know enough about vitamins to ask very specifically about a certain vitamin, but would instead ask a more general question: "Do cereals contain Vitamin B?" And this is where the engines must be intelligent enough to know that the answer to this question might be found in documents about "Thiamine in foodstuffs".

The processing capability makes the difference

"When artificial intelligence first became a possibility in the 1950s and 1960s, the Americans spent large sums of money on research and development. But they discovered that language was a far more complex phenomenon than they imagined. And for this reason most of the developments came to nothing," says Troels Andreasen, head of the Intelligent Systems Laboratory and project manager for the OntoQuery project in the Department of Computer Science at Roskilde University.

"But today, search engines today have such powerful processing capabilities that they open up completely new possibilities for carrying out the complex calculations needed to teach machines to understand what we want," he says. For example, statistical methods can be used to decide how and in what connections certain words are used on the internet. It also makes it

possible to equip the engines with knowledge such as an ontology to enable them to reason and gain a better understanding of what we are asking for."

"The idea is not to create one big ontology covering all concepts and all knowledge in the world," he says. There will be a whole range of different ontologies, each connected to specific subject areas, which will work together to add new perspectives to searching the Internet.

Ankiro: When businesses train search engines and chat robots

The development of intelligent, ontology-driven search engines also has commercial perspectives, not only for large search engines such as Google and Yahoo, but also among the smaller software developers.

The software company Ankiro in Copenhagen started their hunt for intelligence already in 1998. Despite some typical IT turbulence at the end of the nineties, the company today has 14 employees and can prove that search engines *can* be taught things to improve their performance. The company has developed engines such as Jubii for large Danish organisations including the national association for town and city councils (Kommunernes Landsforening) and the international financial consultancy KPMG.

Not words but concepts

"We started out buying comprehensive dictionaries of Danish words and synonyms to build a database of over 60,000 words with their grammar and morphology," explains Bo Vincents, the director of the company. "But for a search to be intelligent, it cannot just be based on a *word* such as 'economy', but on the actual *concept* of 'economics', incorporating all inflections of the word, and all synonyms and all super- and subordinate terms related to this core meaning of 'economics', as well as all the synonyms of these terms with their super- and subordinate terms and so on in a widening network of conceptual links. These 'secondary' conceptual values (embodied in words) also form part of our search strategy, though they will not be given the same central weight," explains Bo Vincents.

The user as starting point

"Language use has individual coloration, and people often use words that are not 'official' or 'standard'. So our task is to take the actual terminology of the user as our starting point when we train the search engines," he says. "If a user is looking for information about environmental services on the town and city councils' web page, he might type in "garbage collection " in the search box, even though this is not the 'official' term used in municipal discourse. The engine must be able to make the connection between the user term "garbage collection" and the web page's "environmental services" to deliver an appropriate response.

Furthermore, many users have problems with spelling words or make typing errors when inputting. So Ankiro has equipped its search engine with a phonetic spell checker, which will recognise and correct typing and spelling errors.

Chat robots serve users

This same ontology-based technology is also used in Ankiro's chat robots, such as *Rosa*, which answers questions about the party's politics on the web page of the Danish Social Democratic Party [LINK: www.socialdemokratiet.dk], even if the questions are completely mis-spelt.

As well as dictionaries and ontologies, Rosa and Ankiro's other chat robots also need what is called a *parser*, which uses language rules to analyse the grammatical form of an input

sentence to find the right answer to the user's query. "The parser is trained to work out which combinations of e.g. subjects, sentence types and tone should lead to which answers. This way the dialogue robot can give an appropriate answer – even if the user uses a rather intimate or rude style of speech," Bo Vincents explains.

Speech will drive natural language technology

One of the great challenges for the researchers at Ankiro is to connect the search engines with the chat robots, so that the users can choose whether they want to interact in natural language with the system or simply use mainstream keywords when looking for information on the web.

"Today, it is difficult to see the perspectives for communicating with the search engines in a natural language," Bo Vincents admits. "But this technology will really come into its own relevance when Danish speech technology is ready to go mainstream. Then we will *talk* to the computer, and it will be unnatural for us to communicate in anything but natural language."

[End of article]

About the OntoQuery project

OntyQuery is a Danish, interdisciplinary project between The Centre for Language Technology, The Technical University of Denmark, Copenhagen Business School, Roskilde University and The University of Southern Denmark.
Homepage: www.ontoquery.dk
E-mail: het.id@cbs.dk

About the author

Maria Bernbom is a business graduate and a journalist. She is the director of the communication agency Reflekt, and apart from her work in copywriting, consultancy and teaching, she writes articles about language and communication as well as other subjects.
Homepage: www.reflekt.dk
E-mail: maria@reflekt.dk