

# Natural Language Processing for intranet searches

---

## Introduction

The purpose of this research document is to explore how and if Natural Language Processing (NLP) should be used for searching on intranets (as opposed to using NLP for searching on Internet sites). The question that must be answered in particular is: for intranets, does the cost of implementing NLP outweigh the benefits?

---

## Overview of NL searches

One application of Natural Language Processing (NLP) is to use let users employ everyday language to search databases. Most common database search engines in use today use Boolean searching (*cat and dog and man and bit*), leaving the connection between these subjects undefined. NLP search engines can be queried with the likes of 'find me the man that bit the cat and the dog', which signifies the relationship between the subjects and leads to more accurate searching.

*Note:* NLP and speech recognition are often perceived as similar technologies. Indeed, both technologies use pattern recognition to enhance the human-computer dialogue. But that is where the similarity stops. Actually, in a one possible scenario, the results of speech recognition would be fed into a natural-language processor.

---

## Practical uses of NLP

The majority of NLP search engines in use today are on publicly searchable Web sites. The most common is the Web search engine Ask Jeeves (<http://www.ask.com>) but NLP is increasing being used in e-commerce especially for large and complex databases.

Generally, it is being used instead of classic Boolean searching (although many NLP search engines can also handle Boolean searches) to enable a more 'user friendly' facility for users who have neither substantial skills in searching databases nor have a comprehensive knowledge of the format of the database.

To get an impression of how users search for information using NLP, see <http://www.ask.com/docs/peek/>. Do note that what you are seeing here is not what the users really entered but rather how Ask Jeeves rephrased their questions. You can tell that this is the case because the statements are very uniform and contain no spelling errors.

To catch a glimpse of how people search on typical web search engine, see [http://www.metaspj.com/spy/filtered\\_b.html?shadow=96156186.6441](http://www.metaspj.com/spy/filtered_b.html?shadow=96156186.6441)

---

## The limitations of NL searching

NL searching has a few well-known limits:

- Because it is difficult for the users to discover the capabilities of NL searching, they tend to assume that the capabilities are limitless. This makes it difficult to manage users' expectations about NL searches, which causes many users to click away disappointed.
- In order to manage users' expectations, they need to be trained in the proper use of NL searching engines, which defies the purpose of having an 'intuitive' tool.

- NL search engines do not degrade gracefully. Once you reach the limits of what they can do, they may very well perform worse than simple, keyword-based search engines.

---

## Who is using NL searches?

NLP search engines are mainly being used in public Web based environments. This is probably due to the large implementation implications (see below).

The same implementation implications also mean that mostly large companies are implementing NLP search engines. Apart from the Web search engines such as Ask Jeeves, other companies using NLP include Dell, General Motors and Intuit.

---

## How NLP is implemented

Implementation of NLP takes place in two steps:

- Technical set-up of the search engine (software)
- Set-up of the dictionaries and relationships that enable the search engine to use natural language searches. NL search engines rely heavily on this initial step of establishing a list of everyday words (dictionaries) that might be used to search the database and establishing how these words relate to each other (relationships).

NLP is implemented either by purchasing an NL search engine and carrying out the implementation of the dictionaries and relationships internally or by contracting an NLP search engine developer to help in this process.

The advantage of contracting a search engine developer is that they not only provide the software but also considerable expertise in setting up the dictionaries and relationships.

---

## The costs of implementing NLP

NLP is expensive compared to other types of search engines. This is because, unlike other search engines, NLP needs substantial amounts of people resources for the set-up and maintenance of the dictionaries and relationships.

There is not a lot of information about the cost of using an NLP developer however as an indication, the US-based magazine *Industry Standard*, reported *Ask Jeeves* as charging between \$200,000 and \$1 million per year to run a NLP search engine.

---

## Should I implement NLP

The effectiveness of NL searching increases with the domain specialization. For example, it is simpler to implement a effective NL search engine for the 'accounting at company A' than it would be for the broader domains of 'Civil Engineering', 'the 19<sup>th</sup> century' or 'Children of the world'.

Search engines are an extremely important facility for any electronic body of information. On the Internet, users often go straight to the search facility, rather than bother with understanding the navigation principles and categories of a site.

The question here is whether NLP is appropriate for searching within an intranet. It is fair to assume that users of a search engine for an intranet are professionals that have specific questions in a domain with which they are quite familiar already.

That is why we feel that to intranet users:

- NL searching will bring little added value because they can easily be taught the few rules of phrasing effective keyword-based searches
- NL searching may appear a bit overdone and gimmicky, particularly when it does not deliver good results

Our recommendation would be to:

- Focus on adding rich metadata to information
- Focus on creating thesauri, dictionaries, proper name databases, etc. (including synonyms and disused terms), possibly allowing for misspellings through phonetic spelling
- Choose a powerful, traditional, keyword-based search engine and *configure it properly*. Such a search engine would eliminate redundancy in search results and excel in relevance ranking.
- Wait and see whether users have sufficient searching power with keyword-based search engines, and if they don't, move to NL searches.
- Consider more semantically oriented engines such as Topic and Autonomy. For an example of Autonomy at work see <http://idm.internet.com/isearch.html>.

---

## Resources

Infoworld, 31 May 1999

[http://www.infoworld.com/researchtools/search\\_f.html](http://www.infoworld.com/researchtools/search_f.html)

Online, May 1999

<http://www.onlineinc.com/onlinemag/OL1999/feldman5.html>

© 2000 Namahn