



Nynodata's TemaSearch

Resources and Techniques for
Multilingual Search

Overview



- Who are we
 - Language Technology Company in Telemark
 - Several Language Products, including “Nyno”
- Who am I
 - Higher Education in Computer Science
 - Lead Engineer
 - Linguistics background

TemaSearch



- The problem
 - Large number and broad spectrum of search engines in use
 - Whole web search (Google, Microsoft, Yahoo)
 - Single site search (FAST, Lucene, CMS, ...)
 - Desktop/Enterprise search
 - Home-grown website search (there are lots of these!)
 - Typically exact match input of words to documents
 - Some perform stemming, but typically not in Norwegian, not to mention Bokmål/Nynorsk
 - Few use synonyms, spelling variations etc.

TemaSearch (contd.)



- The Solution
 - TemaSearch builds on the established search engine, and supplements the search query with words that have the same or similar meaning.
 - Often produces more relevant results and rarely worsens the results. (We'll see why later...)
 - Technology such that other filters, such as spell checking can be provided.

Single Language Search



- Literal keyword matching does not work. Search query differs from the target document
 - Inflected forms (skole vs. skolene)
 - Spelling variations (utdanning / utdaning)
 - Stylistic variations
 - Closed synonyms
 - Open synonyms

Multi Language Search



- Multilingual environment – all the issues of single language search plus more
 - User base submits search queries in several languages
 - Need to find out which language they are using
 - Target pages/documents span several languages
 - Translate search queries and/or documents to match

Multi Language TemaSearch



- Analysis of probable language(s) search query is written in
 - Lexical analysis of input words combined with statistics
- Disambiguation of homonyms – no need to use all word classes
- Queries resources for alternatives for each base form
- Adds words in the search query as hints to the search engine of likely alternatives
- Inflected output: baseforms only, matching inflections for original word, or all inflections for word
- What types of words to add?

Word Alternatives



- Resources in Bokmål and Nynorsk provide words with same or similar meaning
 - Inflect/stem original words
 - Closed synonyms (stylistic & spelling variations)
 - Open synonyms
 - Translations, plus all the above for translated words
 - Strengthens translation dictionaries by adding spelling variations and synonyms

Example



Søk på HiB.no

⋮

Alternativer for arbeidsledighet høyskole

bokmål [synonymer](#)

arbeidsledighet [ledighet](#) [arbeidsløyse](#) [arbeidsløshet](#) [[alle](#)]

nynorsk [oversetting](#)

arbeidsledighet [arbeidsløyse](#)

høyskole [høgskule](#)

Om [Norsk temasøk](#).

Søkeresultater for **arbeidsledighet høyskole høgskole**

Relevant Suggestions



- Resource base is large – filtering selects most relevant suggestions
 - “Disambiguate” homonyms by using most likely word class
 - Suggestions per input word / total suggestions
 - Internal relevance score for alternatives, particularly translations
 - Input and output language filtering / preference

Language Filtering I



- Scenario – Monolingual audience, multilingual site
 - User base primarily use one language (e.g. Bokmål)
 - Document base over multiple languages
- Action
 - Translate input text into languages of document set
- Filter
 - Input languages: prioritize Bokmål, but check for others
 - Output languages: all
- Output
 - Query with original words, plus translated words

Language Filtering II



- Scenario
 - Multilingual audience, monolingual document base
- Action
 - Replace with translation when input language not the same as document base
- Filter
 - Input languages: all
 - Output languages: document base language
- Output
 - When search query in different language, original words are removed and translations used instead

Multi Language Search



- With multiple output languages, we can combine all alternatives in one search query,
- Or, separate them into multiple searches for each language
 - E.g. display different languages in different areas or tabs on the results page
- Translations can be expanded with inflected forms, synonyms. The query may become “diluted” with different word senses, but...

Why it works I



- Additional words are added as optional suggestions, e.g.
 - Skole → Skole OR Skule
- Approximate translation is good enough. Could just as well be
 - Skole → Skole OR ashdfafhkwy
- Will not worsen search results.
- Search engines use ranking to find the most relevant results. Original query is still present - the baseline – additional words do not make the query worse, and often make it better.

Why it Works II



- Query can become “diluted” with other word senses, possibly changing the meaning of the query
 - Words are semantically incoherent
 - May overgenerate, superfluous, redundant, unnecessary and excessive words.
- But
 - Search ranking picks out the most useful and relevant combinations and downplays the others.

Automatic / Manual Suggestions



- By default, all words are added automatically to the search query.
- End user can choose to add words to the query by hand.
- Useful for synonyms and translations covering many word senses.
- Fallback in cases where search engine ranking is not effective

Resources



- Software is language neutral – resources provide all the language knowledge
- Resources in today's TemaSearch
 - Bokmål and Nynorsk
 - Dictionaries, 150k words, ca. 1.6M inflections
 - Closed synonyms (22000 bm, 57000 nn)
 - Open synonyms (17000 bm, 7000 nn)
 - Translation (130k baseforms)

Additional Languages I



- Supporting additional languages possible by adding resources: E.g. Danish, Swedish, English.
- Present system best for closely related languages
 - Effectiveness of distant languages determined by search engine ranking
 - Word by word translation

Additional Languages II



All resources apart from translation are optional, although the more the better!

- Translation dictionary
- Inflected dictionary
- Word frequency lists
- Synonyms
- Disambiguation rules
 - Can be complex or as simple as prioritizing by word class

Future Directions



- Correlate search queries, result documents and documents chosen by users
- Multi-word expressions
- Language model from search engine index
- Domain specific dictionaries: scientific, medial, legal etc...
- Additional languages



Q & A

mat.mcgowan@nynodata.no
www.nynodata.no