# Finding Relevant Passages using Noun-Noun Compounds: Coherence vs. Proximity

Eduard Hoenkamp
hoenkamp@acm.org

Rob de Groot
rob@gang5b.cx

Nijmegen Institute for Cognition and Information (NICI)
PO Box 9104, Nijmegen, the Netherlands

## Abstract

Intuitively, words forming phrases are a more precise description of content than words as a sequence of keywords. Yet, evidence that phrases would be more effective for information retrieval is inconclusive. This paper isolates a neglected class of phrases, that is abundant in communication, has an established theoretical foundation, and shows promise for an effective expression of the user's information need: the noun-noun compound (*NNC*). In an experiment, a variety of meaningful *NNC*s were used to isolate relevant passages in a large and varied corpus. In a first pass, passages were retrieved based on textual proximity of the words or their semantic peers. A second pass retained only passages containing a syntactically coherent structure equivalent to the original *NNC*. This second pass showed a dramatic increase in precision. Preliminary results show the validity of our intuition about phrases in the special but very productive case of *NNC*s.

## Introduction

Many researchers share the intuition that noun phrases (NP's) are a richer and more precise representation of meaning then keywords: 'horse' and 'race' may be related, but 'horse race' and 'race horse' carry more circumscribed meaning than the words in isolation. Hence, several researchers have suggested that using noun phrases instead of keywords, may improve retrieval effectiveness. Empirical studies about the more relaxed definition of phrases (e.g. that keywords be in the proximity of one another in the text) were not conclusive [3]. But given that people use NP's all the time to identify referents they seem worth paying attention to: A system with a syntactic analyzer could match the query NP to the documents. This becomes especially powerful if the NP is extended with meaning preserving syntactic transformations (e.g. from the NP 'horse race' to 'race for horses').

The present research is part of a project on proactive information filtering [6], in which the information need is represented conceptually rather than linguistically. To generate a query, the representation is translated into a particular type of NP, the noun-noun compound (*NNC*). This paper discusses an experiment where the coherence, which the *NNC* inherits from its conceptual representation, can be used to increase the precision for documents retrieved through proximity matching before they are presented to the user.


SIGIR 2000 7/00 Athens, Greece


## Noun-Noun Compounds

Several researchers [8, 5] have described systems that go from conceptual graphs to a query to search the TREC corpus (using boolean search with proximity). In our information filter, the representation language for the information need is ONTOLINGUA [4]. For an information filter, a query is effective if it locates relevant content, be it rendered as a text passage, a graph, or a picture [7]. Since information filters are meant to find information on the user's behalf, they should be autonomous in constructing a query. Hence, filters may have to express concepts for which they do not have a dictionary entry. This, however, is exactly where people use *NNC*s pervasively. They construct new ones on their feet when needed, and may never say them again (hence the technical term 'Hapax Legomena'). Even people with an otherwise deficient grammar (e.g. second language learners), use *NNC*s correctly. Because of its remarkable flexibility to render underlying meaning representation, we wanted to study the efficacy of *NNC*s for information retrieval and filtering.

Some nouns occur more frequently as part of an *NNC* then

others (e.g. 'leader'), hence are called *productive*. This does not mean they always combine into a meaningful phrase. Table 1 shows that the productivity of the nouns does not determine whether a compound is *interpretable*. Since people produce language so as to be understood, the

| | high-productive | low-productive |
|---|---|---|
| high-interpretable | herring knife<br>swamp worm | pizza cook<br>pub joke |
| low-interpretable | rubber mood<br>mussel bomb | disease spark<br>fountain lance |

Table 1: Productivity of a noun (i.e. how often occurs in a compound) does not predict how interpretable it is in a compound (after [2]).

*NNC*s in documents will be interpretable. So in order to match, an effective query must be interpretable as well. The question then arises: how to construct interpretable queries without human intervention? This is a good point to summarize where the focus of this approach deviates from the related approach of 'query expansion':

- The presented approach starts from a conceptual representation,
- It uses a simple, restricted, yet pervasively used construction to identify referents,
- It incorporates knowledge to only produce phrases that people would actually use.

Fortunately, much is known about the way people process novel nominal compounds [10], and our filter implements rules that govern interpretation of compounds [1] to generate interpretable *NNC*s. We will use the term *coherence* to indicate that two nouns combine to express a concept rather than just be in textual proximity. Now, given that the filter can generate *NNC* queries from a conceptual representation, the next step to measure how effective this construction is in retrieving relevant passages from a corpus.

### The corpus for the experiment

It may be clear that the ideal collection for our experiment would be:

- a corpus that is grammatically tagged, making the experiment independent of the power and correctness of a parser,
- which contains a substantial portion of spoken language, so that novel *NNC*s are likely to occur,
- which contains a variety of subject areas, to insure generality of the results.

Such is the British National Corpus (BNC), a 100 million word collection of samples of written and spoken language from a wide range of sources[1]. The whole collection has been grammatically tagged. We used the 'BNC sampler', a 10% subset of the corpus, where the grammatical tags have been manually checked, and which consists of about an equal amount of spoken and written texts.

### Method

For the experiment, we first selected a set of pairs $N_1N_2$, each pair representing a different way that an $N_1$ and $N_2$ can produce a coherent meaning. For example, the pair 'repair job' represents the case of an $N_2$ of which $N_1$ is a particular instance. The pair 'ice cube' represents the case where $N_2$ indicates that it consists of $N_1$. Adams [1] describes thirteen such pairs, and we made an exemplar for each.

Next, for each such pair $N_1N_2$, WORDNET was used to generate new compounds by replacing either noun with its synonym or hyponym. This set of compounds then, forms a collection of *NNC* queries with similar meaning. For each query in this set, a boolean *proximity matching* isolated potentially relevant passages in the BNC documents. Proximity matching is defined by Alta Vista's NEAR operator (a 10 word window), for comparison with related research [9]. The match was performed over each document in the BNC sample. Each matching passage was scored by hand by two raters, and marked as relevant if the raters agreed. Next the precision was calculated. This was repeated for all sets of compounds.

For *coherence matching* to occur, a passage must contain one of the following;

- an exact match with $N_1N_2$,
- a coherence preserving syntactic transformation, such as 'cube of ice' for 'ice cube',
- a match of the above types where either noun might be replaced by a synonym or hyponym, e.g. 'repair job' replaced by 'maintenance work'.

Coherence matching was performed over the documents already found during proximity matching. The reason is twofold: First, it is a stricter match than proximity matching (which is akin to matching by 'query expansion') so there are no other documents it could match. Second, it models the aim of the present research, to approach the WWW by first gathering documents through a search engine. The passages were scored for precision by two raters, as before.

[1] http://info.ox.ac.uk/bnc/

## Results

The results of the experiment are summarized in table 2. For example, the cell 'proximity' vs. 'relevant passages' gives the total number of relevant passages found for the thirteen types of noun-noun compounds, matched over the whole BNC sample. For each type the precision was calculated, and averaged over all types. This is the number in the last column of table 2. In the case of coherence

| | relevant passages | irrelevant passages | average precision |
|---|---|---|---|
| proximity | 36 | 152 | .43 |
| coherence | 27 | 38 | .72 |

Table 2: Comparison of proximity and coherence matching for the thirteen types of noun-noun compounds, and average precision over type. Confining matches to semantically invariant syntactic transformation of the original *NNC* query shows a significant precision gain from .43 to .72 (Wilcoxon,$N = 13$, one-tailed, $p < .01$ ).

matching, the number of documents is about a third of those scored in proximity matching. Then the question arises: could this increase in precision have occurred by chance? We don't know the distribution of relevant passages in the BNC per query, so we need a non-parametric test for this. Further, coherence matching occurs over a subset of those for proximity matching (as it is a second pass), hence the observations are correlated. Under these conditions the Wilcoxon test is appropriate, a simple non-parametric test for correlated paired observations. At this test the increase is significant, at the level of $p < .01$.

## Conclusion and future work

The work reported here is part of a continued effort to build a proactive information filter where, as in all filters, queries have to be constructed on behalf of the user. Given that the major search engines depend on linguistic input, the user's conceptually represented information need has to be translated into linguistic form. We argued that the *NNC* query is a promising candidate to translate into, and based on this, we defined the 'coherence matching' algorithm. To assess its merits, we had the sample BNC corpus indexed by AltaVista, and modeled a proximity match after AltaVista's NEAR operator. We found that coherence matching produced a dramatic gain in precision over proximity matching.
Coherence matching, therefore, is an important tool to select relevant passages from potentially relevant material received from a search engine, before results are passed on to the user.

## References


[1] Valerie Adams. *An introduction to Moderm English word formation.* London: Longman, 1973.

[2] Riet Coolen. *The semantic processing of isolated novel nominal compounds.* PhD thesis, University of Nijmegen, 1995.

[3] W. Bruce Croft. Effective text retrieval based on combining evidence from the corpus and users. *IEEE Expert*, 10(6):59–63, 1995.

[4] Adam Farquhar, Richard Fikes, Wanda Pratt, and James Rice. Collaborative ontology construction for information integration. Technical report, Knowledge System Laboratory, Stanford, 1997. Retrieved May 1997 from WWW, http://ontolingua.nici.kun.nl/.

[5] David Gardiner, John Riedl, and James Slagle. TREC-3: Experience with conceptual relations in information retrieval. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 333–352. NIST, 1994.

[6] Eduard Hoenkamp. Spotting ontological lacunae through spectrum analysis of retrieved documents. In Asuncion Gomez-Perez and V. Richard Benjamins, editors, *European conference on artificial intelligence ECAI-98, Workshop on ontologies and problem solving methods*, pages 73–77, 1998.

[7] Eduard Hoenkamp, Onno Stegeman, and Lambert Schomaker. Supporting content retrieval from WWW via 'basic level categories'. In Marti Hearst, Fredric Gey, and Richard Tong, editors, *SIGIR '99: 22nd annual international ACM SIGIR Conference*, pages 311–312, aug 1999.

[8] Elizabeth Liddy and Sung Myaeng. Dr-link: A system update for TREC-2. In *Proceedings of the Second Text REtrieval Conference (TREC-2)*, pages 85–100. NIST, 1993.

[9] Dan Moldovan and Rad Mihalcea. Using wordnet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1):34–43, 2000.

[10] Mary Ellen Ryder. *Ordered chaos*, volume 123. University of California press, 1994.