

Merging domain thesauri with a generic Wordnet: A case study with the Dutch Army Museum

Antal van den Bosch
Tilburg University
Antal.vdnBosch@uvt.nl

Abstract

Domain thesauri and Wordnets are complementary structures with different purposes. Yet, they can be connected to offer some extra functionality on both ends; the thesaurus can be enriched with more higher-level hypernyms, potentially offering query expansion and recommendation options in search applications, while on the other hand the Wordnet can be enriched with the domain terms of the thesaurus. A case study was performed on two thesauri of the Dutch Army Museum and the Dutch EuroWordNet. The case study reveals that while the hypernyms in the thesaurus are hardly found in the Wordnet as they refer to domain terms, the Wordnet offers many hypernyms that can be reliably connected to terms in the thesaurus. This constitutes a considerable enrichment, enabling richer query expansion and search recommendations.

1. Introduction

This report documents a case study which set out to investigate the possibility of an automated procedure for enriching two thesauri with hypernymic terms from a generic Wordnet. The two thesauri central to this case study are maintained by the Dutch Army Museum. They both offer lists of basic terms, used as metadata tags in the databases of their collections of military books and objects, respectively. It is felt by the maintainers that although the thesauri suit their purpose, they are relatively isolated because of their bottom-up driven origin; they have not been linked "upwards" to other more generic resources. On the one hand, it is simply a matter of fact that the thesauri cover the domain, while more generic lexical resources such as lexicons, thesauri, or encyclopaediae generally do not cover domains such as the military domain well, or only the most common part of it. On the other hand, it may be useful to link up the thesauri to standard generic thesauri or Wordnets, to be able to do intelligent query expansion or recommendation to users in searching the collections.

The two thesauri, one intended for the physical objects collection, the other for the museum's library, already contain a three-layered structure of basic terms, sometimes linking to narrower terms, and more often linking to hypernyms, i.e. broader terms, which act as the "parent node" to sets of basic terms ("sibling nodes"). The maintainers of the thesauri are interested in the question whether the layer of hypernyms could perhaps be expanded by connecting the thesauri to a more generic thesaurus or subsumption ("is-a-kind-of") hierarchy, the backbone structure of the relational database structure which has become known under the name "Wordnet" by the work of Miller and colleagues at Princeton (Miller et al., 1990).

The case study described in this paper describes how we measured the overlap between the two thesauri and a Wordnet, namely the Dutch part of EuroWordNet,

quantitatively. As expected, the two thesauri do not overlap completely with the Wordnet; rather, they contain many terms the Wordnet does not have. In the other direction, we note a substantial number of basic terms in the thesauri that also exist in the Wordnet, thereby offering links to a hyponym and a full hyponymic path through Wordnet. A qualitative analysis shows that this latter overlap can be exploited in a fully automated enrichment procedure, linking 24% and 40% of the basic terms in the two thesauri, respectively, to the Wordnet.

The report also provides an overview of observations that may be relevant and may be a cause for manual annotation, correction, and enrichment, but that do not merit automatic enrichment procedures. We end the report by summarizing our recommendations.

2. Term statistics of the thesauri and the Wordnet

The Dutch Army Museum currently uses separate thesauri for its museum and library objects. In the case study, the two thesauri have been considered independently. As Table 1 shows, the “KG” thesaurus for library objects is the larger of the two. Both thesauri offer, on top of a list of basic terms, a set of hypernyms. For query expansion purposes, it is optimal if a hypernym is connected to more than one basic term, so that queries on one of the terms may be expanded by the sibling terms having the same hypernym. Yet, with both thesauri about 40% of the hypernyms are only connected to a single basic term. Although these may be used for query expansion themselves, they do not offer the possibility to expand to sibling terms.

<i>Thesaurus</i>	<i>Number of basic terms</i>	<i>Number of hypernyms</i>	<i>One-time hypernyms</i>
KG	38,067	2,692	1,118
PM	11,337	384	156

Table 1. Term and hypernym statistics of the two Army Museum thesauri.

The Wordnet used in the study is the Dutch extension of EuroWordNet, a multilingual database with “Wordnets” for several Dutch, Italian, Spanish, German, French, Czech and Estonian (Vossen, 1998). The Wordnets are structured in the same way as the Princeton WordNet for English (Miller et al., 1990): they are connected graphs of synsets (sets of synonymous and near-synonymous words) with basic semantic relations between them. As hyponym and hypernym relations are by far the most frequent semantic relations in Wordnet and EuroWordNet, it is helpful to see a Wordnet as a subsumption (“is-a-kind-of”) hierarchy linking narrower terms (hyponyms) to broader terms (hypernyms). Each Wordnet represents a unique language-internal system of lexicalizations. In addition, the Wordnets are linked to an interlingual index, to which also the Princeton Wordnet is linked. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any of the other languages.

The Dutch EuroWordNet contains 44,015 synsets (34,455 nominal synsets, 9,040 verbal synsets, and 520 adjectival/adverbial synsets), and distinguishes 70,201 senses (Vossen et al., 1999). Hypernym and hyponym relations between all terms in the Wordnet determine the major hierarchical structure, although many more less

frequent relations exist in the Wordnet (e.g. antonymic, holonymic-meronymic, agentive, instrumental, causal, etc.). In this study we only consulted the Wordnet for existing terms and their relations to hypernyms through the `ewnpy` tool¹ written by Erwin Marsi.

3. Overlap between the thesauri and the Wordnet

We measure both the overlap and the lack of overlap between terms in the thesauri and the Wordnet. A lack of overlap is to be expected with specialized domains with relatively many jargon terms. Overlap indicates that the domain contains common words. The latter may not be unproblematic as overlapping words may actually have different senses in the two resources. We begin with a quantitative analysis of the overlap, followed by a qualitative analysis in order to establish a ground for recommendations on the automatic enrichment of the thesauri.

3.1 Quantitative analysis

Both thesauri contain hypernyms that do not occur as terms anywhere in the Wordnet. In hierarchical terms, they have upward pointing links that do not connect with the Wordnet hierarchy. As expected, these terms denote typical domain words, i.e. military terms, and must be assumed to be useful for searching in the domain. On the other hand, the Wordnet contains a substantial number of terms that also occur as basic terms in the two thesauri, and therefore the Wordnet, due to its complete hierarchical structure, naturally offers hypernyms for all these terms.

Thesaurus	Links to hypernyms		
	Thesaurus	Both	Wordnet
KG basic terms	9,749	2,553	5,635
PM basic terms	1,647	285	1,367

Table 2 . Numbers of links from basic terms in the two thesauri to hypernyms in the thesaurus and to Wordnet, and the number of basic terms that link to hypernyms in both resources.

Quantitatively, as shown in Table 2, it can be observed that a considerable number of basic terms in both thesauri can be matched to hypernyms in the Wordnet. To take the example of the larger KG thesaurus, the Wordnet offers hypernyms of 5,635 basic terms. Of this set of matching terms, 2,553 already have a hypernym according to the thesaurus. Thus, the number of unique hypernyms that could be imported from Wordnet is $5,635 - 2,553 = 3,082$, which would constitute an increase of about 24%. For the smaller PM database, this increase would be close to 40%. In other words, Wordnet could be used for substantial enrichments of the hypernym layer in the thesaurus.

It can also be observed in Table 2 that there is not that much overlap in the hypernyms already in the thesaurus, and those found in the Wordnet. With the

¹ <http://ilk.uvt.nl/~mars/software/ewnpy.html>

smaller PM thesaurus, the overlap consists of a mere 285 terms having hypernyms both in the thesaurus and in Wordnet; seen from the thesaurus, this is only 15%; in the case of the KG thesaurus, this relates to 21%.

3.2 Qualitative analysis

The quantitative overlap statistics do not provide insight into the usability of the overlap. To arrive at an automatic procedure for enriching the thesauri with Wordnet hypernyms, we first turn our attention to the basic terms in the thesauri that have hypernyms both in the thesaurus and in the Wordnet (i.e., the middle statistic in Table 2). Through a manual analysis of randomly picked cases with both thesauri, we established that in the vast majority of cases in which a basic term links to hypernyms in the thesaurus and in Wordnet, the two hypernyms are different. In the minority of cases in which they are the same, no enrichment is possible nor needed, as the Wordnet agrees on the hypernym already present. The question is, rather, whether the Wordnet's hypernyms are in any way more appropriate than the thesauri's hypernyms. Our analysis shows that the thesauri offer hypernyms with the proper military sense, while the Wordnet offers an alternative sense that is typically not military. An example is the term "klaverblad" (clover leaf) which according to the thesaurus is kind of "schouderbedekking" (shoulder coverage), and which the Wordnet lists as a kind of "blad" (leaf) or "verkeersplein" (traffic crossing point). Clearly, the thesaurus hypernym should be kept. Introducing the Wordnet hypernyms would contaminate the thesaurus with non-military senses that users will probably not be intending to use. From this analysis we conclude that it is not advisable to introduce any of the Wordnet hypernyms to the basic terms which already have a hypernym in the thesaurus.

Probably the most interesting category of cases is represented in the right-most column of Table 2: the cases in which a basic term has no thesaurus hypernym, but in fact it is present in the Wordnet, and thus has a hypernym. We already pointed to the substantial number of cases in this category for both thesauri. These cases could all be processed in an automatic enrichment procedure that would add 24% and 40% more hypernyms in the KG and PM thesaurus, respectively. A qualitative analysis of a randomly selected number of cases reveals that indeed, automatic enrichment is merited; all hyponyms are appropriate and deemed useful for further tasks such as search query expansion, even in cases where the Wordnet hypernym appears a little too general.

Moreover, the qualitative analysis revealed the surprising fact that the Dutch version of EuroWordNet contains many military terms and hypernyms that the Army Museum thesauri do not contain yet. For example, "deportatie" (deportation) is a kind of "verbanning" (banishment) or "straf" (punishment) according to the Wordnet; a "handgranaat" (hand grenade) is a kind of "granaat" (grenade). The Wordnet has stored that a howitzer is a cannon, that a caliber is either a size, a measure, or a level, and that a jeep is a terrain vehicle.

3.3 Other observations

The qualitative analyses described in the previous subsection caused us to observe several other minor phenomena in the thesauri that merit attention, but that do not lend themselves easily to automatic enrichment procedures.

First, the thesauri contain internal consistency errors that are revealed through analysis such as the ones performed. Some hypernyms make mentions of hyponyms, i.e. basic terms, which are actually not present as basic terms (372 cases in the KG thesaurus, and 67 cases in the PM thesaurus). Also, some hyponym-hypernym relations are accidentally reversed. These cannot be easily detected, but linking the terms to the Wordnet may reveal that their relation in the Wordnet is reversed; this semi-automated correction method would arguably be more feasible than manually checking the full thesauri. Finally, we have spotted several misspellings, causing potentially significant misses in query expansion when the misspelling occurs in a hypernym.

Second, the Wordnet offers the concept of synsets, viz. a set of words with synonymous or near-synonymous meaning, which are typically sibling words under the same hypernym. Both basic terms and hypernyms in the two thesauri have synset words. Table 3 lists the synset statistics of the thesauri. Roughly, one in every ten basic terms has a synset term, while about one in every three thesaurus hypernyms has a synset term in Wordnet. Qualitative analysis of these synsets reveal that synset terms often do not relate to the military domain, disallowing automated enrichment. Also, while the basic terms already present have a bottom-up reason to be in the thesaurus, new synonymous words do not; it may only be of help in limited cases to have synonymous hypernyms to assist users with expanding or refining queries.

Thesaurus	Number of basic terms	Synset words	Number of hypernyms	Synset words
KG	38,067	3,637	2,692	888
PM	11,337	1,367	384	108

Table 3. Synset statistics of the thesauri against the Wordnet.

Third, the hypernyms in the thesauri can also be looked up in the Wordnet to see if they in turn have a hypernym, possibly allowing the addition of a second layer to the thesauri. Indeed, in the larger KG thesaurus, 1,235 hypernyms of the 2,692 have yet another hypernym in the Wordnet; in the PM thesaurus, 153 of the 384 hypernyms have a Wordnet hypernym. It would be possible to automatically enrich the thesauri with these second-layer hypernyms, but many may be too general terms to be useful.

Fourth, although we have simply maintained the fact that the two thesauri must exist independently, we did compute their overlap. It turns out that the two thesauri do not share a lot of their terms. About 80% of the basic terms (9,022 terms) in the PM thesaurus do not occur in the KG thesaurus, while about 94% of the basic terms in the larger KG thesaurus do not occur in the PM thesaurus. In terms of hypernyms, 53% of the PM hypernyms do not occur in KG, while about 95% of the hypernyms in KG do not occur in PM. Although the smaller thesaurus still shares a reasonable portion of its terms with the larger one, they clearly aim at different subdomains, i.e. collections of physical objects versus library objects; these results certainly do not merit recommending automatically merging the two thesauri, which would only be merited if they already overlapped to a large extent.

4. Recommendations

On the basis of our quantitative and qualitative analyses, we recommend that the two thesauri of the Army Museum be automatically enriched by linking to the Dutch EuroWordNet. All basic terms in the two thesauri that currently do not have a "broader term" (hypernym), but that do occur in the Wordnet, should be linked to the Wordnet. The link can be limited to simply adding the single hypernym from the Wordnet to the thesaurus, but the thesaurus may also simply link to the Wordnet, offering full access to the Wordnet's rich relational structure (i.e. containing of hypernymic-hyponymic relations, but also many other types of relations). We considered other types of automatic enrichment based on overlaps and non-overlaps, but found none to be as reliable as the one described above.

In general, we recommend that, after the thesauri have been enriched as recommended, search engines running on both thesauri are developed that exploit the expanded hypernymic-hyponymic structure to the fullest. Query expansion should be tested, and also explicit ways of guiding searches (widening, narrowing down), and recommending "related" topics by traversing the hypernymic-hyponymic relation to find parent and sibling nodes to terms currently being used in search.

Acknowledgements

This study was financed as part of the RNA project (<http://www.rnaproject.org/>). The author gratefully acknowledges Erwin Marsi (for the ewnpy tool), Annet Ruseler, Maud Kornaat, Thijs Jonkman, Hans Nederbragt, Sander van der Meulen, Steven Mije, and Piroska Lendvai for advice and help.

References

Miller, G., R. Beckwith, C. Fellbaum, D. Gross and K. Miller (1990). *Five Papers on WordNet*. CSL Report 43. Cognitive Science Laboratory. Princeton University.

Vossen, P. (Ed.) (1998), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.

Vossen, P., Bloksma, L, and Boersma, P. (1999). *The Dutch Wordnet: Version 2, Final*. ILLC Technical Report, University of Amsterdam.