Computational Linguistics in the Netherlands Journal 4 (2014) 17-28 Submitted 05/2014; Published 12/2014

Activating Qualified Thesaurus Terms for Automatic Indexing with taxonomy-based WSD

Inga Kohlhof* I.KOHLHOF@ELSEVIER.COM
Boris Kozlov* B.KOZLOV@ELSEVIER.COM
Marius Doornenbal* M.DOORNENBAL@ELSEVIER.COM

Elsevier B.V., Amsterdam, The Netherlands

Abstract

Many thesauri contain a number of descriptors consisting of the term proper plus a suffix in brackets meant to explain the term's intended interpretation. For instance, the MeSH thesaurus contains a term *Polymorphism (Genetics)*. For different thesauri, these terms account for 1%-5% of all descriptors. For automatic indexing based on recognizing term occurrences in free text, these terms are practically useless —free text never or very rarely contains term references of this form. A naive text annotation method, matching these terms with their bracketed qualifiers stripped off (the 'bare' terms) results in frequently wrong interpretations. We investigated to what extent short forms of qualified terms (viz. *Polymorphism*) can be disambiguated by looking for concepts in their textual environment that are ontologically related to the represented concepts (in casu, *Genetic Polymorphism*), or to the concepts used to qualify (*Genetics*).

Using the NLP framework of the Elsevier Fingerprint Engine® we created a set-up to test disambiguation for a set of 30 qualified terms from the NAL thesaurus, that we annotated in approximately 1500 scientific abstracts from the agricultural domain found in Scopus®. By their ambiguity with respect to the NAL Thesaurus we distinguished three groups of test terms: Terms with unqualified homonyms, terms with qualified homonyms and terms without homonyms inside the thesaurus. For all three groups, the best results (65-75% recall, 83-93% precision) are found when both the concept hosting the qualified terms and the qualifier concept are used to identify supporting concepts in the terms' contexts. Like similar Word Sense Disambiguation (WSD) techniques our approach is attractive as the system is informed by existing knowledge and therefore does not require huge knowledge-intensive investments. At the same time the system delivers reasonable precision. For these reasons we will seek to refine it to bring up recall scores.

1. Introduction

Many thesauri contain a number of descriptors consisting of the term proper plus a suffix in brackets meant to explain the intended interpretation of the term. Examples are in (1).

(1) a. UV-B (ultraviolet radiation) (NALT)
b. files (tools) (NASA)
c. Vital Energy (Philosophy) (MeSH)

The ISO's[1] standard for monolingual thesauri for information retrieval, ISO 25964-1, recommends to use qualifiers to distinguish otherwise homographic terms within a thesaurus:

> "Homographs (sometimes referred to by the broader term 'homonyms') are words with the same spelling but different meanings. [...] When homographs are needed as thesaurus terms, the meaning of each term should be clarified and the traditional way to do this is by adding to it a qualifier in parentheses. The qualifier should be as brief as possible, ideally consisting of one word. Often a broader term, the qualifier should

1. International Organization for Standards

indicate the context or subject area to which the concept belongs. It forms part of the term and does not serve as a scope note." (ISO 25964-1, 6.2.2)

In a sample of 6 thesauri the ratios of these qualified terms (QT) range from 0,6% to 5,7% (see table 1).

Thesaurus	Domain	# of QT	# of terms	% of QT
Geobase	Geosciences	67	11,146	0.6%
Gesis	Sociology	96	11,443	0.8%
MeSH	Medicine	8,261	739,923	1.1%
NAL	Agriculture	1,527	87,17	1.8%
Compendex	Engineering	1,072	21,576	5.0%
NASA	Astronomy and Physics	129	22,771	5.7%

Table 1: Ratios of qualified terms in thesauri (see Thesaurus References)

For automatic indexing, these terms are practically lost: When left intact they will, unless the qualified term and its qualifier occur in close proximity, not be found; when simply stripped of their qualifiers, the remaining ('bare') terms are likely to be ambiguous. We investigated to what extent qualified terms (with the qualifier term removed, 'bare' terms) can be disambiguated during text annotation. We pursued a disambiguation solution based on looking for concepts in the textual environment which are ontologically related to the concepts that the bare terms represent and/or to the concepts that the qualifiers represent. In this WSD task a word sense corresponds to a thesaurus concept with a unique concept identifier. The tested technique is knowledge-intensive but unsupervised WSD in the sense of Navigli (2009), meaning that the technique does not use training data and supervised machine learning techniques but relies on 'knowledge' represented in external resources, in our case, a thesaurus. As our test thesaurus we chose the US National Agricultural Library (NAL)'s thesaurus (NALT for short). Its ratio of qualified terms is moderate (see above), but the thesaurus is large and thoroughly maintained, with a dense net of relations between its concepts.

Our framework is the Elsevier Fingerprint Engine®, a Natural Language Processing (NLP) suite primarily used to compute concept vector representations —referred to as 'fingerprints' —of scientific abstracts in Elsevier's Scopus® database. The Elsevier Fingerprint Engine[2] is a suite of NLP functionalities comparable to frameworks such as GATE[3] or UIMA[4]. The current application of the framework includes the continuous semantic indexation of documents of all scientific domains with seven thesauri and two vocabularies. The resulting concept representations are further used in applications for different purposes, for instance author profile presentation, cf. Vestdam et al. (2014).

In the context of this framework we are developing a number of WSD techniques. Specifically for the annotation of qualified NALT concepts in agricultural scientific abstracts, we look for ontologically related concepts of the concepts that are represented by ambiguous terms in the NAL 'fingerprints' provided by the Fingerprint Engine. The relationship between the occurrence of one (unambiguous) concept in a text that supports the annotation of another (ambiguous) concept in the same text, we termed *licensing*.

A good example of the issue is the ambiguity between *cones (retina)*, being part of the eye anatomy, and *cones (plant)* that are seed-cones. The former term is licensed by the occurrence of, for instance, *visual spectra*, whereas seed-cones may be licensed by a word such as *flower*.

2. http://www.elsevier.com/online-tools/research-intelligence/products-and-services/elsevier-fingerprint-engine
3. https://gate.ac.uk/
4. https://uima.apache.org/

2. Related work

There is a rich diversity of approaches to the problem of Word Sense Disambiguation in Automatic Language Processing. The review paper by Navigli (2009) still provides a very helpful survey of the field. Within the family of knowledge-based techniques, which generally "compare the context of the ambiguous word to the information available in a terminological resource" (Jimeno-Yepes and Aronson 2012, p.42), using general-purpose sources like Roget's International Thesaurus (Roget 1911) or WordNet[5] contrasts with approaches using domain-specific sources like ontologies and thesauri. Faralli and Navigli (2012) observe that WSD "has been oriented towards domain text understanding for several years now."

Ontological / thesaurus relations between concepts have been put to use in different ways; often other sources of information are exploited at the same time. Below we give some examples to illustrate the heterogeneity of approaches and to convey an idea of the range of their performance. The references in the quoted articles can serve as a stepping stone to explore the field.

Alexopoulou et al. (2009) evaluate the performance of three different approaches, two of which make use of relations between concepts of the Gene Ontology (GO) and of the US National Library of Medicine's Unified Medical Language System (UMLS), respectively. Their 'Closest Sense' approach compares the lengths of the shortest paths from all co-occurring concepts to the competing concepts. The 'Term Cooc' method computes co-occurrence graphs not only for (concepts hosting) ambiguous terms and co-occurring GO/MeSH concepts but also for their descendants (the authors refer to this as *Inferred Co-occurrence*). The methods are reported to achieve average accuracies of 77 and 81%, contrasting with 96% achieved with the third tested approach using a training set of high-quality metadata. The authors conclude that where training data is not available "ontologies can play a very important role to improve disambiguation" but that their contribution depends on their size, consistency, density and the degree to which their hierarchical relations are true is-a relations.

Jimeno-Yepes and Aronson (2010) test four different ways of assigning one of several competing concepts to ambiguous terms in the largest biomedical ontology and thesaurus, the UMLS (see above). Each of those four disambiguation methods is based on a comparison of the context of ambiguous terms (title and abstract) to a representation of the competing concepts, making use of their definitions, synonyms, (hierarchically and non-hierarchically) related concepts and semantic types (hypernyms) in the UMLS. For instance, one approach automatically extracts a corpus from MEDLINE, another approach uses MEDLINE's Journal Descriptors. For their test set of 49 ambiguous UMLS terms, Jimeno-Yepes and Aronson (2010) report accuracies between 0.59 and 0.75[6].

Prokofyev et al. (2013) propose a hybrid method that leverages a background knowledge-base as well as corpus statistics. They test their approach with the ScienceWISE ontology for the domain of physics and with the MeSH thesaurus for the biomedical domain against a selection of supervised and unsupervised measures. All tested knowledge-based methods are based on the minimal and the average distance of a concept in the knowledge base to its neighboring concepts and the number of such neighbors. Each of the knowledge-based methods is clearly outperformed by all (MeSH) or at least the best (ScienceWISE) of the tested supervised methods employing Binary Concept Context Vectors. However, the combination of all knowledge-based methods with that latter method scores best.

While ontologies are commonly encoded using ontology languages, i.e. formal languages, the links thesauri maintain between concepts differ considerably with respect to logical strictness and consistency as well as the kinds of supported relations. In addition, thesauri differ with respect to the granularity of the represented concepts and the 'density' of their net of concept relations. Results reported for attempts to exploit thesaurus relations for WSD are therefore hard to compare from the outset. On top of all this, we investigated a specific kind of usually neglected thesaurus terms. Since our method is based on the presence of closely related concepts in the context of ambiguous

5. http://wordnet.princeton.edu/
6. *Accuracy* defined as the number of *Instances Correctly Predicted* divided by the number of *All Instances*

terms which aren't always around we only achieve recall scores of 61-75%. Our results for precision of 86-91%, however, are comparable to those of others, including the ones summarized above.

3. Ambiguity of the qualified terms

According to the ISO standard quoted above qualified terms should be introduced to distinguish homographs within a thesaurus.[7] The first natural question to ask is: How ambiguous are the NALT qualified terms when stripped of their qualifiers?

	Qualified terms (QT) in NAL	Examples	#	%
1	without corresponding terms in NALT	WMV2 (Watermelon mosaic virus) coffee (beverage) solute movement (soil)	699	45.8%
2	with synonymous corresponding terms in NALT	American Indians USED FOR Native Americans (non-Alaskans) Native Americans	37	2.4%
3	with NALT bare term homographs	age determination (trees): age determination Malaya (country): Malaya ← Culicidae USED FOR mosquitoes peduncle (nerves): peduncle ← plant anatomy	236	15.5%
4	with NALT QT homographs	DCPA (chlorthal-dimethyl): DCPA (propanil) kiwis (birds): kiwis (fruit) water level (groundwater): water level (surface water)	600	39.3%
5	with NALT bare term and QT homographs	jerky: jerky (fruit): jerky (vegetable) shipping: shipping (animals): shipping (by air): shipping (by land)	55	3.6%

Table 2: Homographs of qualified terms in the NALT

As Table 2 shows, almost half of the QT have no lookalikes whatsoever in NALT (46%, first row), some 2% have synonymous correspondences, and for slightly more than half of the qualified terms (rows 3-5) competing terms with different meanings exist in the NALT. Row 5 gives the overlap between the QT with bare or QT homographs. In addition, there is overlap between terms with synonymous corresponding terms and all three other groups[8].

So if we simply removed the QTs' qualifiers we would end up with a number of homonyms within the thesaurus. As for the qualified NAL terms without thesaurus-internal competitors: their intended restricted application is marked by their qualifiers. Even a cursory glance at them shows that many compete with homographs not represented in the NALT:

7. Homographs are words which are spelled alike but differ in meaning. Words, in turn, are often, and particularly in automatic indexing, identified with the set of their possible - genitive, plural etc. - forms. During indexing these forms are mapped to their base form by some kind of normalization procedure. So homographs are words with different meanings which are spelled alike, modulo normalization.
8. which is why the percentages do not add up to 100.

(2) a. hunting (animal behavior) evidently competes with hunting behavior of humans,
b. the city of Acre (Brazil) has a homograph in acres of land,
c. grafting (plants) competes with the surgical procedure of grafting.

Assigning the NAL concepts hosting the QT to these competitors would be just as wrong as their assignment to their competitors within NAL.

When looking for NAL concepts in the context of ambiguous terms that 'support' one of its possible meanings we may be able to make use of what qualified terms have that others don't: their qualifiers.

4. The qualifiers

To serve as a stepping stone to identify NAL concepts that support the meaning of an ambiguous term, qualifiers should best be NAL concepts themselves. A high proportion, namely 1,232 out of 1,527 or 81% of the qualifiers of NALT's QT correspond directly to a NAL term. The rest of the qualifiers are either terms not present in NAL (3a), more complex qualifiers, e.g. prepositional phrases containing a NAL term (3b), adjectives (3c), NAL terms distinguished by qualifiers (3d, NADPH is represented by two qualified terms in NAL (NADPH (coenzyme), NADPH (nicotinamide adenine dinucleotide phosphate)), terms with meta-descriptors (3e) or others. We made no efforts to include these QTs in our experiment, and in the following, the expressions 'qualified terms' and 'QT' will refer only to NAL QTs with NAL-referential qualifiers.

(3) a. sex determination (analysis)
b. thinning (of canopy)
c. wood production (biological)
d. glutamate synthase (NADPH)
e. Anacystis nidulans (unspecified)

The following table (Table 3) shows which relations hold between NAL-referential (NR) qualified terms and their qualifiers.

NAL QT with	Example	#	% of NR-QT	avg. dist.	range of dist.
Synonymous qualifiers	jaundice Used For icterus (jaundice)	218	18%	0	0
Ancestral qualifiers	Enhydra (Compositae) IS-A Asteraceae IS-A Compositae	550	44%	1.62	1-5
Non-ancestral qualifiers	emergence (insects)	473	38%	4.94	1-12

Table 3: Semantic relations between qualified terms and their qualifiers

Almost 20% of the QT with NAL-referential qualifiers are qualified with a term that is also an entry term to the concept hosting the qualified term. The term *icterus (jaundice)* is such a case. Unless the qualified term is a subordinate concept, in these cases, the qualified term and its qualifier are synonyms —for instance *Culcita (Asteroidea)*, where *Culcita* is a genus of cushion stars belonging to the class *Asteroidea*. Almost half of the qualified terms have ancestral qualifiers, i.e., their qualifiers sit somewhere above them in the NAL hierarchy. Also, we note that the average distance between qualified term and qualifier term is small. No ancestral qualifier is more than 5 steps away from the qualified term; For 85% of the QT with NAL-referential qualifiers, qualifier terms are no more than two nodes away. The non-ancestral qualifying concepts of the remaining

almost 40% of the QT can be located twice as far away, cf. the range of distances, and their average distance is significantly longer, viz. almost 5 nodes. Non-hierarchical paths are widely distributed: There are 134 different paths, the most frequent one —where the qualifier is a related concept of the qualified term —connecting only about 7% of the term pairs.

5. Test set-up

We picked 30 test terms (i.e. 2.5%) from the set of 1,232 NALT QT with qualifiers that refer to NAL concepts (see section 4 above). With regard to ambiguity these fall into three groups: Half of them (15) do not compete with homonyms in the NAL thesaurus (*hunting (animal behavior)* is such a term), 10 compete with other qualified terms (like *cones (retina)* which competes with *cones (plant)*), and 5 have unqualified competitors in NALT (e.g., the QT *peduncle (nerves)* competes with *peduncles*, a descendant of plant anatomy).

Two groups of QT are not represented in our set of test terms. First, there are 37 NAL QT with synonymous bare terms (cf. section 3). Evidently they do not compete with these. In automatic indexing they are redundant —their synonym will be assigned (unless otherwise restricted) to all occurrences of the term —and we delete them. Second, there is the group of NAL QT with both qualified and unqualified competitors. None of our test terms belongs to this group; we will get back to it in section 7.

The sizes of the three groups of test terms - 15, 10 and 5 terms, respectively - reflect the distribution of these ambiguity classes in the whole set of NALT QT, cf. (3) above. In a second step we annotated all good indexings of these terms in 1,500 Scopus documents (consisting of titles and abstracts) of scientific articles from the agricultural domain with *brat*, NacTeM[9]'s web-based annotation tool[10].

That done, we were able to measure recall and precision of the test terms' indexation in varied configurations. At the same time, we kept track of tokens indexed with two different concepts (homonyms) in an uncurated reference test set of 50,000 documents from the agricultural domain.

We kept the parameters listed in 4 constant throughout our tests.

(4) a. No mutual licensing was allowed, i.e. qualified terms were not allowed to license each other.

b. Only distinct licensors counted, i.e. it did not matter how often a licensing concept occurred in a text.

c. We looked for licensors within a window of 1,000 tokens, 500 on the left and 500 on the right hand side of ambiguous terms. In most cases this comprises the complete document (title and abstract).

d. We tested two sets of licensors, ontologically (i.e. as NAL concepts) close to the concept hosting the base term and to the concept hosting the qualifier concept of each ambiguous term respectively:

i Concepts related to the host concept of the qualified term:
SYN,REL,BT,REL-BT,BT-BT,REL-BT-BT,*NT-BT,NT,NT-NT,NT-NT-NT*

ii Concepts related to the host concept of the qualifier:
SYN,REL,BT,REL-BT,BT-BT,REL-BT-BT,*BT-BT-BT,REL-BT-BT-BT*[11]

Both sets focus on hierarchical relations.

9. http://www.nactem.ac.uk
10. http://brat.nlplab.org
11. where SYN=synonyms, REL=related concepts, NT=parents, NT-BT=siblings, NT-NT = grandparents, BT=children, BT-BT=grandchildren, REL-NT=parents of related concepts, NT-REL=concepts related to parents etc. Relations not shared by the two sets are italicized.

Given that —depending on the logical strictness of the subordinate relations in a thesaurus descendant concepts generally share their ascendants' properties —we assumed that descendants make no worse licensors than their ascendants. The licensing kin of the base term concepts includes parents and grand-parents, children and grand-children while the licensing kin of the qualifier concept includes only descendants, down to the 4th grade (grand-grandchildren). As for related concepts (concepts connected by a 'related term' link), we kept close to the reference concepts, including only first grade related concepts and their descendants to the 2nd grade (grandchildren). Since, as can be seen in table 2, qualified terms and their qualifiers can be very close to one another, licensors defined referring to the base term and licensors defined referring to the qualifier concept will often overlap. The set of licensors generated based on a QT's qualifier concept (4di) will, on average, contain 5 times as many licensors as the set of licensors generated based on the concepts hosting the qualified term (4dii). That is because qualifiers tend to be more general in meaning and found in higher positions in the NAL hierarchies than qualified terms (many hierarchical relations in NAL are class-element relations, and more than 57% of the qualified terms even have scope notes which are direct ancestors or children of direct ancestors); higher-level concepts, in turn, tend to have more descendants (this is logically true of classes and their elements) and more related concepts than lower-level concepts.

6. Baseline scores and lexical filters

To compute baseline scores we stripped all qualified terms of their qualifiers and indexed the annotated documents with the resulting thesaurus. Table 4 shows recall and precision of the indexation of our test terms as well as the number of occurrences of the qualified test terms with homonyms in NAL (indexing homonyms) which in this baseline configuration is equal to the number of occurrences of those terms.

Type of qualified term (QT)	#	Recall	Precision	Homonyms
with regular (bare) NAL homonyms peduncle (nerves): peduncle ← anatomy	5	100%	3.9%	253
with qualified NAL homonyms cones (retina): cones (plant)	10	98.1%[12]	55.5%	1282
without NAL homonyms inheritance (genetics)	15	100%	72.6%	0
				1535

Table 4: Baseline scores for three groups of qualified terms

The baseline precision scores for the three groups of QT differ considerably. The extremely low score of 3.9% for QT with bare NAL homonyms results from the fact that the meanings of the qualified terms are used much more rarely (in the test texts, from the agricultural domain) than the meanings of the bare NAL competing terms, e.g. *peduncle* is very often used for plant peduncles, much less often for peduncles of nerves. In contrast, the 55.5% precision for qualified terms with qualified competitors suggest that their meanings are generally more equally distributed. The precision measured for the indexing of qualified terms without homonyms in NAL justifies the skepticism about their actual ambiguity expressed in section 3: >30% of the term's occurrences in our test texts do not have the meaning represented by the qualified term's host concept. A precision of just above 70% is clearly below the precision of 80% (that we defined as the minimum target

12. The NAL thesaurus was updated after manual annotation; for that reason, a few occurrences of our test terms were not indexed any longer when these tests took place.

precision for thesaurus concept identification, for use in our automatic indexing processes), and strikingly below the average precision of NAL terms of almost 95%.

The analysis of the baseline performance shows that removing qualifiers without further measures does not suffice to activate qualified terms for automatic indexing because for none of the three groups of qualified terms, not even the ones without competitors in NAL, we achieved acceptable precision scores.

Next, we defined additional filters for test terms reflecting properties represented in their lexical entries. For instance, we let tokens be indexed with names likes *Acre* (the city) or *Anemia* (the fern) only if their first letter is a capital, the term *gums* as a representative of the concept *gingiva* must be met in plural form, etc. As shown in the following table, these filters brought up indexing precision for our test terms by almost 5% while reducing the number of terms indexed with more than one concept by almost 4.4%. All subsequent tests were performed applying these filters.

Type of qualified term (QT)	#	Precision with lex. filters	Precision Δ	Ind. homs. with lex. filters	Ind. homs. Δ
with regular (bare) NAL homonyms peduncle (nerves): peduncle ← anatomy	5	6.9%	+3%	121	66
with qualified NAL homonyms cones (retina): cones (plant)	10	55.4%	-0.1%	1303	0
without NAL homonyms inheritance (genetics)	15	81.3%	+8.7%	0	0
			+4.8%	1424	66 =4,4%

Table 5: Indexing homonyms for three groups of qualified terms before and after application of lexical filters

7. Successful configurations

We achieved the best results for all three groups of test terms when we used the united sets of licensors defined based on the concept hosting the qualified term (the base term concept) and licensors defined referring to the concept hosting the term's qualifier. The precision of at least 80% that we require from productive thesaurus terms (cf. section 6) defines a lower boundary of acceptability. For two groups, QT with qualified competitors and QT without competitors in the NALT, a single licensor per abstract warranted acceptable precision scores; for the third group, the terms with unqualified competitors in NALT, a second licensor was necessary and sufficient (cf. table 6).

As expected, requiring more licensing concepts in the context reduced recall rates — not, however, for the first group of QT with non-qualified NAL homonyms. This group consists of only 5 rare test terms; all their occurrences happen to be supported by 2 licensors. Still, as expected, requiring more licensors increased precision rates — but only for the first and the third group of test terms. Additional licensors in the contexts of test terms of the second group, QT with qualified NAL homonyms, did not help to disambiguate them.

Table 7 shows how the two sets of licensors contribute to these scores. When qualified terms have competitors in the NALT (groups 1 and 2) the correct interpretation of terms can be identified more reliably, i.e. more precisely, by looking for kin of the concept hosting the qualified term (the base term concept) in the context (4^{th} column). Given that the kin of the base term concept is generally much smaller than that of the qualifier concept (cf. section 5), unfortunately but not

Type of qualified term (QT)	#	R/P with a min. # of licensors of	
		1	2
with regular (bare) NAL homonyms peduncle (nerves): peduncle ← anatomy	5	70.0% / 77.8%	70.0% / 87.5%
with qualified NAL homonyms cones (retina): cones (plant)	10	65.0% / 83.2%	30.9% / 83.1%
without NAL homonyms inheritance (genetics)	15	75.1% / 93.1%	57.4% / 95.7%

Table 6: Most successful configurations for three groups of qualified terms

Type of qualified term (QT)	#	Best results using both sets (cf. table 6)	Results (R/P) using kin only of the	
			base term concept	qualifier concept
with regular (bare) NAL homonyms peduncle (nerves): peduncle ← anatomy	5	70.0% / 87.5% (min. 2 licensors)	30% / 100.0%	60.0% / 85.7
with qualified NAL homonyms cones (retina): cones (plant)	10	65.0% / 83.2% (min. 1 licensor)	32.5% / 90.5%	51.4% / 81.5%
without NAL homonyms inheritance (genetics)	15	75.1% / 93.1% (min. 1 licensor)	53.8% / 92.0%	53.8% / 94.8%

Table 7: Results using kin of the base term / qualifier concepts only

surprisingly, clearly less qualified terms with competitors can be identified on this basis. Quite in contrast, qualified terms without NALT competitors were found equally well with the two sets of licensors (though the exactly identical recall scores are a coincidence) and slightly more precisely with kin of the qualifier concept. We have no explanation for this contrast.

Considering all terms were inactive (never annotated on any text) at the outset, our efforts are a major step forward from the original situation. As shown, if we annotate with the stripped qualified terms (peduncle (nerves) <peduncle) baseline scores vary between groups in an interesting way (section 6) - the combinations of P/R scores (table 6) also differ characteristically between the groups.

Indexing homonyms.

A subset of 145 (i.e. 9.8%) of the baseline of 1,486 test term tokens are assigned more than one index with the most successful configurations. We do not allow for ambiguous indexing results to be produced by the Fingerprint Engine. For this reason, we implemented a complementary fallback mechanism for homonyms created by the licensing process. That mechanism inspects the licensing quality of competing terms by checking, in this order, the number of licensors present for each of them, the path distances - across the concept structure in the thesaurus - between licensing and licensed concepts and the distance in the text under inspection between a term's licensor(s) and the term itself. Should these measures be inconclusive, the ambiguous token is assigned the term which is most frequently (unambiguously) licensed in our test set of agricultural documents (cf. section 5).

All homonyms remaining after licensing belong to the second group of QT with qualified competing terms in NALT (table 6, 2nd row). After application of our resolution procedures recall drops to 61.1% (-3.9%) but precision goes up to 85.9% (+2.7%) for this group. For the whole set of test terms recall drops from 71.2 to 69.8% (-1.4%) while precision rises from 89.8 to 90.8% (+1%).

A balanced solution had to be found for qualified terms with both qualified and unqualified competitors in NAL (like *jerky (fruit)*) which competes both with *jerky* —dried meat —and *jerky (vegetable)*, cf. section 3). Since there are only about 50 of these terms we could determine the best performing configuration by comparing indexing results for all instances. The best results are achieved when licensors were defined based on both the base term and the qualifier concepts and only one licensor was required.

With the test terms in all three groups being indexed with a precision of clearly above 80% we decided to trust our licensing algorithm and replaced 1,232 NAL QT by bare terms for which the algorithm automatically set configuration parameters depending on their group membership.

For our reference test set of 50,000 test documents from the agro-biological domain (cf. section 5) indexed with the updated NAL Thesaurus we measured a slightly improved text coverage ratio (22.21% vs. 21.46% without activation of QTs) and a slight increase in term distinctiveness expressed as a concept distribution over documents (Gini-Coefficient moving to 56.29% vs. 56.77% without activation of QTs).

8. Outlook

Our primary objective performing the tests reported here was to preliminarily answer a question: Can licensing, i.e. requiring ontologically related concepts to be present in the context of an ambiguous term, be used to activate ambiguous thesaurus terms explicitly disambiguated by qualifiers for productive automatic indexing? The answer to that question is yes.

We worked with two constant sets of licensors (based on the qualified term's host concepts and on its qualifier's concept, respectively) and we looked for licensing concepts in a constant text window (of 500 tokens left and right of ambiguous terms). Especially with the set of licensors being, in principle, an open set we do not consider trying to systematically test all possible combinations of all settable parameters. Rather, we will design subsequent experiments for two purposes.

1. A wider range of licensors. We were surprised about the disambiguating force of a single licensor within the range of a typical abstract: For the two largest of the three main groups of qualified terms one licensor was enough to achieve precision rates of 83% and 93%. However when requiring just one more licensor we saw recall drop steeply, by 34% and 18% for the largest 2 groups of tested QT. As a consequence what we will test next is to stepwise enhance the set of licensors by including more descendants, terms sharing a qualified term's qualifier, indirectly related terms etc. to see if and to what extent we can improve recall without jeopardizing precision.

2. Licensing non-qualified ambiguous terms. While we are now able to activate unused qualified terms in the thesauri, the overarching aim is to find a working approach for all ambiguous thesaurus terms. Most ambiguous terms are not openly marked as ambiguous by qualifiers, but compete with concepts not represented in their thesauri. An illustrative example is the term *decline* which represents a plant disease in the NAL thesaurus and is clearly competing with other meanings. Hence, the term 'decline' as an NAL Thesaurus term will often be indexed incorrectly in texts from the agricultural domain. All thesauri contain hundreds or even thousands of such terms. If we could, using our test environment, carve out a set of kin and/or related concepts which can be counted on to be present and reliable supporters (as relations should be) to concepts with ambiguous entry terms we may be able to use that same set to license the assignment of thesaurus concepts with unmarked ambiguous entry terms.

References

Alexopoulou, Dimitra, Bill Andreopoulos, Heiko Dietze, Andreas Doms, Fabien Gandon, Jörg Hakenberg, Khaled Khelif, Michael Schroeder, and Thomas Wächter (2009), Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy., *BMC bioinformatics* **10**, pp. 28. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663782/.

Faralli, Stefano and Roberto Navigli (2012), A new minimally-supervised framework for domain word sense disambiguation, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1411–1422. http://dl.acm.org/citation.cfm?id=2390948.2391109.

Jimeno-Yepes, Antonio J. and Alan R. Aronson (2010), Knowledge-based biomedical word sense disambiguation: comparison of approaches., *BMC bioinformatics* **11** (1), pp. 569, BioMed Central Ltd.

Jimeno-Yepes, Antonio J. and Alan R. Aronson (2012), Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation, *in* Luo, Gang, Jiming Liu, and Christopher C. Yang, editors, *IHI*, ACM, pp. 733–736.

Navigli, Roberto (2009), Word sense disambiguation, *ACM Computing Surveys* **41** (2), pp. 1–69. http://portal.acm.org/citation.cfm?doid=1459352.1459355.

Prokofyev, Roman, Gianluca Demartini, Alexey Boyarsky, Oleg Ruchayskiy, and Philippe Cudré-Mauroux (2013), Ontology-based word sense disambiguation for scientific literature, *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, Springer-Verlag, Berlin, Heidelberg, pp. 594–605. http://dx.doi.org/10.1007/978-3-642-36973-5_50.

Roget, Peter Mark (1911), *Roget's International Thesaurus*, 1st ed., Cromwell, New York, NY.

Vestdam, Thomas "Voldemort", Henrik Steen "Saruman" Rasmussen, and Marius "Sidious" Doornenbal (2014), Black magic meta data-get a glimpse behind the scene, *Procedia Computer Science* **33**, pp. 239–244. http://www.sciencedirect.com/science/article/pii/S187705091400828X.

Thesaurus References

Compendex Thesaurus
http://www.elsevier.com/online-tools/engineering-village/contentdatabase-overview

Geobase Thesaurus
http://www.elsevier.com/online-tools/engineering-village/contentdatabase-overview

Gesis Thesaurus
http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/

MeSH
National Library of Medicine, Medical Subject Headings (MeSH).
http://www.nlm.nih.gov/mesh/2014

NAL Thesaurus
National Agriculture Library.
http://agclass.nal.usda.gov

NASA Thesaurus
http://www.sti.nasa.gov/sti-tools

ScienceWISE ontology
http://www.sciencewise.info/ontology