Augmenting Domain-Specific Thesauri with Knowledge from Wikipedia

Olena Medelyan David Milne
Department of Computer Science, University of Waikato
Private Bag 3105, Hamilton, New Zealand
+64 7 838 4021
{olena, dnk2}@cs.waikato.ac.nz

ABSTRACT
We propose a new method for extending a domain-specific thesaurus with valuable information from Wikipedia. The main obstacle is to disambiguate thesaurus concepts to correct Wikipedia articles. Given the concept name, we first identify candidate mappings by analyzing article titles, their redirects and disambiguation pages. Then, for each candidate, we compute a link-based similarity score to all mappings of context terms related to this concept. The article with the highest score is then used to augment the thesaurus concept. It is the source for the extended gloss, explaining the concept's meaning, synonymous expressions that can be used as additional non-descriptors in the thesaurus, translations of the concept into other languages, and new domain-relevant concepts.

Categories and Subject Descriptors
D.3.3 [**Artificial Intelligence**]: Natural Language Processing – *text analysis.*

General Terms
Algorithms, Experimentation

Keywords
Thesauri, Wikipedia, word sense disambiguation

1. INTRODUCTION
Domain specific thesauri describe the terminology and semantics of a particular field. They have a wide range of applications, from manual indexing and browsing to automated natural language processing. Creating these structures manually is a demanding task that we have so far failed to automate. Thesauri rely on the input of a small group of contributors and thus are often restricted to narrow domains. They are prone to gaps in terminology and bias in topic coverage, especially as domains evolve.

In this paper we introduce a new method of supplementing domain-specific thesauri with information extracted from Wikipedia. This massive online repository of knowledge offers many opportunities for augmenting human efforts to make thesauri stretch further, customize them for new resources or tasks, to fill gaps and maintain them, add desirable new features such as glosses or multilingualism. The core challenge is to identify correct articles for ambiguous thesaurus entries. We tackle this by using contextual terms defined by the thesaurus and contextual articles provided by Wikipedia.

Sections 2 and 3 describe the mapping challenge: the former analyses the specific difficulties faced when mapping concepts from one knowledge resource to another, while the latter presents our solutions. In the second part of the paper we apply our technique to the agricultural thesaurus Agrovoc.[1] Section 4 demonstrates that the accuracy of the algorithm comes close to that of humans. Section 5 investigates and quantifies the contributions that Wikipedia can make to the thesaurus once the mappings are identified. We conclude the paper with related work on disambiguation to Wikipedia articles, and summarize our contributions in this work.

2. COMMON MAPPING PROBLEMS
The first step in augmenting a thesaurus with Wikipedia is to map concepts defined in one resource to the relevant concepts in the other. This process presents several challenges: synonymy and alternative expressions, because there are different ways of referring to the same concept; ambiguity, because the same term might have different meanings; and missing information that might lead to no mappings or erroneous ones.

2.1 Synonymy and alternative expressions
Human language has a tendency to use different terms to explain the same concepts. For example, a very young child might be referred to as an *infant*, *toddler*, or *baby*. Furthermore, almost every organism has a scientific name, a different common name, and may have several regional colloquialisms, e.g. *beetles* are scientifically referred as *Coleoptera*.

Most thesauri have so-called 'non-descriptors' which contain alternative terms for concepts and point to the preferred ones, called 'descriptors'. Similarly, Wikipedia has 'redirects', which contain alternative URLs and titles: a different way of getting to the same article. The thesaurus's alternative terms and Wikipedia's redirects increase the odds of matching concepts between the two structures.

There are many other language phenomena beside synonymy that cause term variations, such as equivalent spellings (*color* vs. *colour*), plurals (*baby* vs. *babies*), abbreviations (*N.Y.C.* vs. *New York City*), equivalent syntactic constructions (*predatory birds* vs. *birds of prey*), and alternative word types (*social economics* vs. *social economies*). Some of these difficulties are addressed by the resources themselves, but manually generated thesauri do not rigorously document every minor variation. Fortunately there are many automatic term conflation methods available, such as stemming, stop-word removal, application of

[1] http://www.fao.org/agrovoc

abbreviations and alternative spelling dictionaries, and word order normalization. These must be applied carefully, however, as even subtle changes indicate significant differences in meaning: *communes* are not the same as *communities* just as *bird cages* are different from *caged birds*. Stronger term conflation strategies yield higher recall (more possible mappings), but lower precision (more incorrect mappings).

2.2 Polysemy

Polysemy or ambiguity of meaning is another challenge in concept mapping. Take the example of *virus*. A thesaurus in the medical domain would only consider biological viruses, while Wikipedia, which attempts to describe all domains, contains several other possible senses, including viruses that infect computers, as well as some proper names. For each thesaurus term, one must collect all the potentially valid Wikipedia articles, and then choose the best.

Wikipedia documents ambiguity with disambiguation pages which list the possible meanings along with short scope notes to explain them. Additionally, specifications in brackets (e.g. *Virus (band)* for a band from Norway) can be stripped out when looking for relevant articles.

To assist in identifying the correct article, Wikipedia contains extensive descriptive information about each candidate which can be compared to context found in the thesaurus (i.e. glosses or related concepts). Context-independent statistics, such as the number of incoming links or positions in disambiguation pages, can also be used to identify the most probable meaning.

2.3 Missing information

Further problems are caused either by missing information in the resources or deficiencies in their structure. Obviously, a concept cannot be found if it is not described in Wikipedia at all or if the equivalent expression for the same concept is missing (e.g. Wikipedia contains an article for *notopterus* which is only accessible through *notopteridae*). It is more difficult to identify cases where the correct meaning of a descriptor is missing. For example, there are a number of Wikipedia articles about *callisto*, but none of them describes the butterfly genus used in our thesaurus. Such cases will inevitably cause missing or erroneous mappings.

Often two concepts in a thesaurus are absorbed in one Wikipedia article (*breadmaking* and *bread*), or one thesaurus concept is expanded into several Wikipedia articles, e.g. *harness* is split into *horse harness*, *safety harness* and *dog harness*. This phenomenon is called 'meaning conflation' and requires some heuristics based on compromises to resolve.

3. THE MAPPING ALGORITHM

We have designed an algorithm to maximize coverage of thesaurus terms and the accuracy of their mapping to Wikipedia, while overcoming problems described in the previous section. We first describe how, for each thesaurus concept, we collect all possible Wikipedia articles, and then show how thesaurus and Wikipedia knowledge about the concepts is used to disambiguate the correct meaning. To keep the explanations clear we use pseudo-code of the algorithm (Figure 1) and explain the semantic relatedness measure applied in the disambiguation process separately.

```
1   extendThesaurus (thesaurus T, wikipedia W) {
2       foreach concept C in T {
3           p = descriptor of C
4           R = set terms related to C
5
6           Wa = getBestMapping(p,R)
7
8           // retrieve information about Wikipedia article Wa
9           // and add it to the concept p
10      }
11  }
12
13  getBestMapping (term p, related terms R)  {
14      M = getAllMappings(p)
15
16      case 1 |M| = 0   return null
17      case 2 |M| = 1   return M.first
18      case 3 |M| > 1   {
19
20          S = set of support Articles
21          foreach related term r in R
22              S = S U getMostCommon(getAllMappings(r))
23
24          if (|S| = 0)
25              return getMostCommon(M)
26
27          M = weightMappings(M,R,S)
28          return M.best
29      }
30  }
31
32  getAllMappings(term t) {
33      L = list of Wikipedia names whose titles match t
34      foreach l in L {
35          if (l is a redirect)
36              l = target of l
37          if (l is a disambiguation page) {
38              D = getDisambiguationTargets(p,l)
39              L = L U D
40          }
41      }
42      return L
43  }
44
```

Figure 1. Pseudo-code for extending a thesaurus with information from Wikipedia

3.1 Collecting candidate mappings

For all thesaurus terms and names of all Wikipedia pages, we first strip all explanations in brackets and apply stemming to identify a list of matching titles (Figure 1, line 33). To deal with term variations (Section 2.1), we only use the first stage of Porter's Stemmer, which removes the plural endings *-s*, *-ess* and *-ies*. More aggressive conflation strategies, such as full stemming, removing stopwords and word-reordering introduce too many inaccuracies. For stronger variations such as abbreviations, equivalent expressions and synonyms, we rely on Wikipedia redirects, which are very accurate [4]. We do not use non-descriptors in Agrovoc as synonyms because they serve other purposes too.

In the next step, we check the type of the encountered Wikipedia page. If it is a redirect, we follow its target (Figure 1, line 35). If it is a disambiguation page, we analyze its content to identify more candidates (Figure 1, lines 38 to 40). Disambiguation pages in Wikipedia are not written in a

consistent way, but the following heuristics help retrieve relevant meanings fairly accurately:

- include the first link on the page;
- include the first link in each list item explaining each meaning, unless the search term appears in the explanation as plain text;
- ignore all links in the section 'see also'.

3.2 Resolving ambiguity

The next algorithm step is to choose the correct article from the candidate set, if it has more than one member (Figure 1, lines 18 to 29). We use related thesaurus entries as context to identify the intended sense (Figure 1, line 4). We consider as 'related' the following: non-descriptors or alternative labels for the given term, as well as descriptors of broader, narrower and sister concepts of this term. In the *virus* example, this context might be broader terms such as *micro-organism*, related terms such as *pathogen* and *bacteria*, and narrower terms such as specific viruses. For sparse thesauri, we can crawl further to identify more context terms—siblings, grandparents, grandchildren—if necessary. We take the top 10 concepts with at least one mapping to any Wikipedia article and use them as our context for disambiguation.

We resolve the ambiguity by identifying the candidate article that most strongly relates to the same context. First the relevant Wikipedia article for each context term is retrieved. Again ambiguity is an issue: e.g. one of the context terms for *virus, bacteria*, is another name for rabbit programs—a little known form of malicious software. Considering this sense in the disambiguation process would increase the chances of picking the wrong meaning. Thus, if for a context term more than one mapping is possible, we pick the most common one, quantified by the highest number of incoming links (Figure 1, lines 20 to 22).

Semantic relatedness is computed between each candidate article and all context articles as described in Section 2.3. The best article can then be chosen as the one with the highest average similarity across all contextual articles (Figure 1, lines 27 to 28).

3.3 Computing semantic relatedness

In previous work [5], we have experimented with semantic relatedness, which quantifies the strength of relations between different concepts. For example, one might say that *cash* and *currency* are 95% related, or *currency* and *bank* are 85% related. Despite the evident subjectivity, people are capable of fairly consistent judgments. For example, in [2], 13 participants individually defined relatedness for 350 term pairs and achieved an average correlation of 79% between each individual's judgments and those of the group.

The measure used here quantifies the strength of the relation between two Wikipedia articles by comparing the articles that link to them. This is modeled after the Normalized Google Distance [1], which considers term occurrences on web-pages rather than link occurrences in Wikipedia pages. Formally, the measure is:

$$sr(a,b) = \frac{\max(\log f(a), \log f(b)) - \log f(a,b)}{\log M - \min(\log f(a), \log f(b))},$$

where *a* and *b* are the two articles of interest; *f(a)*, *f(b)* and *f(a,b)* are the number of articles that link to a, b, or both respectively; and M is the total number of articles in Wikipedia. This yields a correlation of 72% with the above mentioned manual judgments when the 353 term pairs used in [2] are manually disambiguated to the appropriate articles.

4. EVALUATION OF THE MATCHING ALGORITHM

This section investigates the performance of our matching algorithm by applying it to the Agrovoc thesaurus. This domain specific thesaurus was created by the U.N. Food and Agriculture Organization to help organize its vast repository of reports. We considered it to be a good test of our techniques because of its relatively large scale—some 28,000 descriptors and 11,000 non-descriptors—and high degree of specificity. The version of Wikipedia we matched it to was released on November 11, 2006. At that point it contained approximately two million articles and a further million redirects.

4.1 Defining the gold standard

To evaluate the accuracy of the disambiguation technique we manually constructed a gold standard with 400 thesaurus concepts and their corresponding Wikipedia articles. The sample of terms is random, but manipulated to evenly cover different levels of ambiguity.

This was done by first automatically identifying the ambiguity of all terms in Agrovoc with respect to Wikipedia—that is, the number of candidate articles they could possibly resolve to according to the candidate selection process described in Section 3.1, cf. Figure1. Concepts with zero mappings were not considered. The remainder were grouped into four bins: those with one mapping; 2 or 3 mappings; 4-10 mappings; and greater than 10 mappings. A hundred concepts were randomly sampled from each bin, then disambiguated by taking the context provided by Agrovoc and manually browsing Wikipedia for the appropriate article. In each case the authors had to agree on the correct mapping.

26 terms were discarded because Wikipedia did not contain the intended sense. For example, Wikipedia has several articles for *Scylla*, but none for the genus of crab described in Agrovoc. Another complication is that a direct one-to-one mapping between concepts was not always possible (Section 2.3). For instance, Wikipedia describes variants of *Ottis* (ear inflammation) in separate articles, while Agrovoc conflates them to a single term. All of the articles are correct, but choosing one discards other important aspects of the concept. This occurred ten times within our gold standard, and was addressed by considering any of the sub-topics to be correct.

The final result was 374 manually disambiguated Agrovoc terms, with a minimum of 89 terms in each bin. To prevent over-fitting, this was kept separate from a similar test set of 200 terms used during development to tweak the algorithm.

4.2 Baseline

We have additionally created a baseline algorithm to investigate the importance of context when disambiguating terms. It operates without mining any additional information from the thesaurus, by selecting the most obvious, well-known sense for each term. Obscure senses can often be ignored by simply choosing the article whose title matches the term exactly: e.g. *Mother* vs. *Mother Nature* or *Mother (song).* If there is no such article, stemming is applied and expressions within brackets are removed, resulting in a list of candidate pages. When searching for *yams*, this results in *Yam (vegetable), Yam (god), Yam (route)* and the disambiguation page *Yam*. Disambiguation pages are not expanded upon but rather replaced with the first

Table 1. Accuracy of disambiguation algorithms (%)

mappings	1	2 or 3	4 to 10	over 10	average
baseline	100	90	78	79	87
our method	100	93	88	88	92

article they link to. The final article—*Yam (vegetable)*—is then chosen as the one which is most referenced by other Wikipedia articles.

4.3 Results

Table 1 summarizes the accuracy of both proposed methods: the baseline and the full algorithm described in Section 3. Each value expresses the percentage of correct mappings within the different bins of ambiguity defined in the gold standard.

The baseline performs surprisingly well. One would expect that the highly specific Agrovoc might use obscure senses of terms—e.g. *Zeus* as the scientific term for *John Dory* rather than the name of a god—but this was largely not the case. For at least ¾ of terms the most obvious sense was the best one. However, considering candidates whose titles do not match Agrovoc's descriptors and using semantic similarity to contextual terms to differentiate between them improves accuracy by up to 10 percentage points depending on the level of ambiguity involved. This translates to a significant reduction of errors: 25% for terms with one or two mappings, and around 40-45% for more ambiguous ones.

Figure 2 shows what occurred when our algorithm ran over the entire thesaurus. Approximately 13,000 terms—or 40% of Agrovoc—was covered by Wikipedia. The vast majority of covered topics were unambiguous, with only 16% matching to multiple articles. This gives an overall predicted accuracy (weighted to reflect the number of terms for each level of ambiguity) of 98% for our algorithm and 97% for the baseline. Of course the reason for this seemingly tiny improvement is that the vast majority of terms have only 1 mapping.

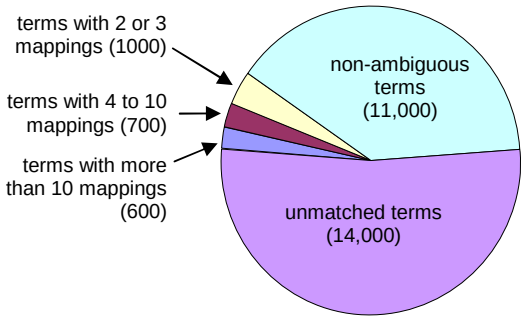

Figure 2. Wikipedia's coverage of terms in Agrovoc

4.4 Error analysis and observations

A total of 26 errors were made by the matching algorithm within the gold standard sample. Some were caused by inconsistency in Agrovoc's usage of non-descriptors. For example, the non-descriptor *foot* is a synonym for *feet*, but *blood*'s non-descriptor *Hematology* is an entirely different topic. In the former case the non-descriptor is the name of the correct candidate article, while in the latter it identifies a contextual one. Our algorithm makes errors because it always assumes the latter case, which is more common.

Some failures were due to articles lacking sufficient incoming links to identify accurate measures of relatedness. For example, *bottling* of produce should be matched to the article *bottling line*, but this is barely referenced within Wikipedia. The topic instead matches to the throwing of bottles at concerts. In other cases accurate measures can be calculated, but are not enough to differentiate between closely related concepts, e.g. *pickle* and *pickled cucumber* or *Taiwan* and *Taiwan Province*. Such cases are similar to the meaning conflation effects described in Section 3.1, but here one of the senses can be considered as more correct than the other, and thus we have counted them as errors.

Finally, Wikipedia missed some redirects for syntactically close structures that could not be conflated with stemming alone. This reduces the number of topics mapped, but also introduces inaccuracies: e.g. *deadwood* was mapped to *Deadwood, South Dakota* instead of the article on *dead wood*.

5. AUGMENTING THESAURI WITH ENCYCLOPEDIC KNOWLEDGE

Once thesaurus concepts have been accurately matched to Wikipedia, many opportunities arise for expanding upon them. With the overlap between the structures serving as a starting point, Wikipedia can be explored for new terms and concepts that might address gaps in the thesaurus or extend it to encompass new topics. Wikipedia can also supply desirable new features such as glosses and translations. This section describes these contributions individually, and investigates their utility by applying them to the Agrovoc thesaurus.

5.1 Increased terminology

Wikipedia is edited by a wide community of contributors, with different backgrounds, interests, and levels of expertise. This diversity allows it to capture wide variations in the terms people use to refer to the same concepts. Traditional thesauri do not have the same properties, because they are edited by few individuals. Often they do not even attempt to define a complete vocabulary, but rather focus on the professional's view of the 'ideal' terminology for the domain. In previous work (Milne et al, 2006) we have shown that redirect relations in Wikipedia match almost perfectly the equivalence relations in a domain-specific thesaurus. Thus, by adding redirects as new non-descriptors to the thesaurus, we improve its terminology coverage without compromising its quality.

Wikipedia was able to extend more than 9000 Agrovoc topics by adding more than 18,000 new terms. For example, the average user would not recognize the term *tetraodontidae* unless they consulted Wikipedia to find the more common variants *puffer-fish*, *swell-fish* and *balloon-fish*. Around 1500 expanded concepts were similar cases where Agrovoc unhelpfully references an organism exclusively by its scientific (Latin) name. Other new terms arose from variations in syntactic construction and word choice (*radioactive pollutants* was expanded to *radioactive waste* and *nuclear residue*). Variations in spelling were also apparent; imaginative Wikipedians demonstrated 16 different ways to spell *cockroach* and 9 ways to spell *Minnesota*. While these seem unlikely candidates for inclusion in a thesaurus, they are potentially useful in assisting novice users to arrive at the correct terminology. The same rationale applies to colloquialisms: the plant *datura stramonium* or *thorn apple* sounds innocuous enough, but Wikipedia's synonyms *devil snare*, *crazy tea*, *beelzebub twinkie*, and *zomby cucumber* hint at its hallucinogenic properties.

5.2 Increased topic and relation coverage

Wikipedia's two million articles and the 300 million links between them represent a vast pool of topics and semantic relations. Consulting it can only increase one's odds of covering all topics and relations within a given domain. The challenge is to identify which articles are relevant to the domain, and which links are strong enough to constitute a valid relation between them. We have not yet attempted to automate this process, and consequently are unable to provide meaningful statistics. However, we can provide some compelling cases to highlight Wikipedia's potential for addressing gaps in both topic and relation coverage.

For example, why does Agrovoc discuss *banking*, *loans*, and *credit*, but not *money* or *currency*? How did the indexers decide that four specific deserts (the *Kalahari*, *Thar*, *Gobi*, and *Sahara*) warranted inclusion while all the others—around 90 according to Wikipedia—did not? In each case, the missing topics could have been identified automatically by their strong relatedness to Wikipedia articles found within Agrovoc. Gaps are equally apparent in the thesaurus' relations: why is there no connection between *farm equipment* and *tractors*, or *milk yielding animals* and *milk*? In these cases the missing relation is indicated by a strong semantic relatedness measure between the corresponding Wikipedia articles.

5.3 Automated maintenance

Wikipedia, which received around 3-4 million edits a month in 2006, excels in covering new developments and discoveries. With such rapid growth and response to change, we can expect it to lend great assistance to the tedious task of maintaining thesauri as their domains evolve. This can be considered as a subset of the task described in the previous section, only here we are specifically interested in capturing new topics that arise.

To evaluate whether Wikipedia could lend assistance to maintaining thesauri, we compared Agrovoc releases from 2001 and 2006. 11,000 new non-descriptors or topics were added between the two versions. Of these new topics, 29% existed in Wikipedia and thus could have been introduced automatically. This in itself is a reasonable contribution, but perhaps not the best demonstration of the method. Agrovoc covers a well established domain and thus the vast majority of new topics are highly specific. Almost all the topics Wikipedia failed to cover were obscure species of flora and fauna which are unlikely to be of interest to anyone other than domain experts. We expect much greater contributions for maintenance within more popular, swiftly evolving domains such as politics, entertainment, or technology.

5.4 Scope notes and glosses

Manual creation of explanatory texts such as scope notes and glosses is time-consuming. Consequently most thesauri explain only a small proportion of concepts (usually the ambiguous ones). Only 4% of Agrovoc's concepts have scope notes. In contrast, all the 13,000 Wikipedia articles relevant to Agrovoc contain explanatory text. By convention the first paragraph of each article can be extracted as a description of the topic. If the FAO decided to add glosses to Agrovoc, our techniques would get them halfway there: every topic that was found in Wikipedia has a succinct description available to define it.

5.5 Multilingualism

Currently, there are over 200 different language versions of Wikipedia. 15 versions have over 100,000 articles, and 75 have at least 10,000. Our approach is language independent: it does not use any language specific resources apart from the input thesaurus itself. With the same method we could extend thesauri in any source language that is sufficiently covered in Wikipedia.

Additionally, Wikipedia can be used to translate thesauri into different languages. Whenever the same concept is described in different versions of Wikipedia, cross-links are maintained to allow navigation between them. After mapping a thesaurus entry to the relevant article in the root language of the thesaurus, we can easily translate it into new languages by following its links to other available versions.

This is an area for which, unlike many other thesauri, Agrovoc requires little assistance. Over the years, the FAO has translated the thesaurus into eight languages (Spanish, Japanese, French, Chinese, Arabic, Portuguese, Slovak and Thai) and more languages are still to come. However, Wikipedia could have greatly expedited this translation process: Table 2 shows the 28 languages for which at least 1000 automatic translations are available. There are a further 45 languages with at least 100 translations.

Table 2. Top 28 non-English Wikipedias and the number of Agrovoc concepts they contain.

rank	language	concepts	rank	language	concepts
1	German	6426	15	Norwegian	2364
2	French	6103	16	Hebrew	2038
3	Dutch	4762	17	Czech	1929
4	Spanish	4639	18	Esperanto	1765
5	Portuguese	4627	19	Turkish	1594
6	Polish	4473	20	Catalan	1585
7	Japanese	3860	21	Ukrainian	1251
8	Italian	3544	22	Bulgarian	1245
9	Swedish	3281	23	Indonesian	1178
10	Russian	2851	24	Hungarian	1172
11	Chinese	2850	25	Vietnamese	1125
12	Finish	2738	26	Serbian	1095
13	Italian	2716	27	Slovak	1033
14	Danish	2492	28	Korean	1000

6. RELATED WORK

The main purpose of this paper is to show to what degree manually-created thesauri can be extended with new knowledge mined from Wikipedia. A similar task has been posed by Ruiz-Casada et al. [6]. They iterate over the entries of the Simple English Wikipedia and map them to corresponding synsets in WordNet. For each article they look up whether its title appears in one of the WordNet synsets. If multiple synsets are possible, they compute similarity between the article and each synset. They use the dot product and the cosine measure applied to the article vector and to the synset gloss extended with related words. Our approach differs by using related terms as context, rather than gloss text, which makes it applicable to the majority of thesauri that do not provide glosses.

Note that the size of the Simple English Wikipedia at the time of writing that paper was just 1841 articles, and consequently only 1200 were mapped to WordNet synsets. Evaluation on two samples with 180 cases each has shown that the accuracy is 98% for monosemous and 84% for polysemous examples. Our task is much more complex, with mapping 28,000 of concepts to over 2M articles in Wikipedia, and we still get higher accuracy.

This is the only work to our knowledge in mapping a thesaurus to Wikipedia. Much more research, as shown below, has been

Existing entry in Agrovoc	Wikipedia's contributions
Broader term: Integument	**Categories:** Human anatomy, Sensory organs, Integumentary system, Dermatology
Related terms: Epithelum, Hair **Sister terms:** Shell, Animal cuticle Skin glands, Sebaceous glands, Sweat gland	**New related terms:** Scar tissue, Hair follicle, Hypodermis, Apocrine glands, …
Used for: Callus (animal skin), dermis, animal epidermis.	**Redirects:** Animal skin, Skin cell, Skin type, Cutaneous, Corneocyte, Cutaneous fold, Oily skin, Skin care, Skin disorders
Scope Note: Of animals, for plants use 'Epidermis'	**Gloss:** In zootomy and dermatology, **skin** is the largest organ of the integumentary system made up of multiple layers of epithelial tissues that guard underlying muscles and organs. [1] Skin pigmentation (*see:* human skin color or coloring) varies among populations, and **skin type** can range from dry skin to **oily skin**. …
Translated into: 16 languages	**Translated into: 49** languages

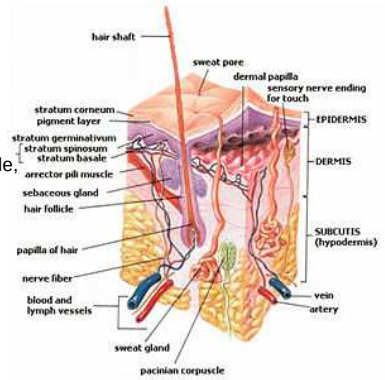

Figure 3. Contributions mined from Wikipedia for the Agrovoc term *skin*.

done in the core problem of this task: mapping terms to the corresponding Wikipedia pages.

Strube and Ponzetto [7] propose the following simplified disambiguating approach. With the WikiMedia software that is a part of the Wikipedia website, for a given term they retrieve the most likely page it can be mapped to. If this is a disambiguation page, they collect all terms and phrases that appear on the internal links on this page. They do the same for the context term and compare the two lists. The link whose title matches one of the terms in the context list is selected, or, if no match is observed, the first meaning listed on the disambiguation page is chosen. Strube and Ponzetto suggest that their solution might not be accurate but it is practical, as they ignore most of the possible meanings listed in disambiguation pages. It is similar to our baseline approach, which is also biased towards the most common sense of a term.

Gabrilovich and Markovich [3] disambiguate terms to Wikipedia articles given a text fragment in which these terms occur. This context could be a simple phrase containing the term in question and its context term (e.g. *Bank of America* to disambiguate *bank*). They first map each word appearing in this fragment to a set of Wikipedia articles in which this term appears. Then a centroid-based classifier is used to rank these articles by their relevance to the text fragment.

Wang et al. [8] propose two methods of disambiguating terms to Wikipedia articles. In the first approach, they compute the cosine similarity between the article of each possible mapping and the document where this term appears. The most similar one is chosen as final mapping. The second method proposed by Wang et al. is restricted to the sentence in which the query term appears. They calculate the conceptual distance for each possible article for the given term to articles of other non-polysemous terms in the sentence. The meaning with the minimum average distance in the Wikipedia category tree is chosen as the result. It is the length of the shortest path between the categories of the articles divided by the depth of the tree.

The main problem with this approach is that the sentence in which the term appears might not have other non-polysemous concepts. The authors do not discuss such cases. Also, it seems unreasonable to discard the information contained in polysemous context terms, as we have shown there are ways of taking them into account. Furthermore, it is not clear which category is chosen if the article belongs to more than one category.

All these methods tackle the problem of disambiguation as part of a larger task. The authors provide little discussion about the encountered difficulties in disambiguation. The problems must have been substantial since Wikipedia's size is immense, and the number of senses per term is much higher than in conventional thesauri. Also, none of the reported experiments are evaluated by the authors intrinsically with a set of manually disambiguated examples. This paper closes a gap in this research area by demonstrating the difficulties of the disambiguation task and evaluating two proposed approaches intrinsically with a gold standard set of mappings.

7. SUMMARY

This paper demonstrates how concepts in a domain-specific thesaurus can be linked to the correct concept descriptions in Wikipedia and what information can be gained from these mappings. Let's look at an example concept to see what we have achieved. Figure 3 compares the original Agrovoc entry *skin* to the new information that we automatically can retrieve from Wikipedia with the proposed techniques. Only new information is shown here, and some of the entries can be further expanded. For example, we included only some of the new terms that Agrovoc is currently missing. This example shows how the efforts of a few humans who have created the thesaurus can be automatically augmented with the contributions of thousands of volunteers, retaining the authority of the former while adding the scale, scope, currency, impartiality and multilingualism of the world's largest example of public collaboration.

8. REFERENCES

[1] Cilibrasi, R.L. and Vitanyi, P.M.B. (2007) The Google Similarity Distance, IEEE Trans. Knowledge and Data Engineering, 19:3, 370-383.

[2] Finkelstein, L., Gabrilovich, Y.M., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2002) Placing search in context: The concept revisited. ACM TOIS 20(1).

[3] Gabrilovich, E. and Markovich, S. (2007) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Proc. of IJCAI'07.

[4] Milne, D., Medelyan, O. and Witten, I. H. (2006). Mining Domain-Specific Thesauri from Wikipedia: A case study. Proc. of the International Conference on Web Intelligence (IEEE/WIC/ACM WI'06), Hong Kong.

[5] Milne, D. (2007). Computing Semantic Relatedness using Wikipedia Link Structure. *Proc. of* NZ CSRSC'07.

[6] Ruiz-Casado, M., Alfonseca, E., Castells, P. (2005) Automatic assignment of Wikipedia Encyclopedic Entries to WordNet synsets, Proc. of Advances in Web Intelligence, pp. 380-386.

[7] Strube, M. and Ponzetto, S. P. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. Proc. of AAAI '06, pp.1419-1424.

[8] Wang, P. and Chen, L. (2007) Improving Text Classification by Using Encyclopedia Knowledge. Proc. of ICDM'07.