Proceedings of the Globalex Workshop on Linked Lexicography, pages 53–60
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Towards a Swedish Roget-Style Thesaurus for NLP

Niklas Zechner, Lars Borin
Språkbanken Text/Department of Swedish
University of Gothenburg, Sweden
{niklas.zechner, lars.borin}@gu.se

Abstract

Bring's thesaurus (Bring) is a Swedish counterpart of Roget, and its digitized version could make a valuable language resource for use in many and diverse natural language processing (NLP) applications. From the literature we know that Roget-style thesauruses and wordnets have complementary strengths in this context, so both kinds of lexical-semantic resource are good to have. However, Bring was published in 1930, and its lexical items are in the form of lemma–POS pairings. In order to be useful in our NLP systems, polysemous lexical items need to be disambiguated, and a large amount of modern vocabulary must be added in the proper places in Bring. The work presented here describes experiments aiming at automating these two tasks, at least in part, where we use the structure of an existing Swedish semantic lexicon – Saldo – both for disambiguation of ambiguous Bring entries and for addition of new entries to Bring.

Keywords: lexicon, word sense disambiguation, topic detection

1. Introduction[1]

1.1. Lexical Semantic Resources for NLP

Lexical-semantic knowledge sources are a stock item in the language technologist's toolbox, having proved their practical worth in many and diverse natural language processing (NLP) applications.

Although lexical semantics and the closely related field of lexical typology have long been large and well-researched branches of linguistics (see, e.g., Cruse 1986; Goddard 2001; Murphy 2003; Vanhove 2008), the lexical-semantic knowledge source of choice for NLP applications is WordNet (Fellbaum, 1998b), a resource which arguably has been built largely in isolation from the linguistic mainstream and which thus is somewhat disconnected from it.

However, the English-language Princeton WordNet (PWN) and most wordnets for other languages are freely available, often broad-coverage lexical resources, which goes a long way toward explaining their popularity and wide usage in NLP as due at least in part to a kind of streetlight effect.

For this reason, we should also explore other kinds of lexical-semantic resources as components in NLP applications. This is easier said than done, however. The PWN is a manually built resource, and efforts aiming at automatic creation of similar resources for other languages on the basis of PWN, such as Universal WordNet (de Melo and Weikum, 2009) or BabelNet (Navigli and Ponzetto, 2012), although certainly useful and laudable, by their very nature will simply reproduce the WordNet structure, although for a different language or languages. Of course, the same goes for the respectable number of manually constructed wordnets for other languages.[2]

1.2. Roget's *Thesaurus* and NLP

While wordnets completely dominate the NLP field, outside it the most well-known lexical-semantic resource for English is without doubt Roget's *Thesaurus* (also alternately referred to as "Roget" below; Roget 1852; Hüllen 2004), which appeared in its first edition in 1852 and has since been published in a large number of editions all over the English-speaking world. Although – perhaps unjustifiedly – not as well-known in NLP as the PWN, the digital version of Roget offers a valuable complement to PWN (Jarmasz and Szpakowicz, 2004), which has seen a fair amount of use in NLP (e.g., Morris and Hirst 1991; Jobbins and Evett 1995; Jobbins and Evett 1998; Wilks 1998; Kennedy and Szpakowicz 2008).

There are indications in the literature that Roget-style thesauruses can provide an alternative source of lexical-semantic information, which can be used both to attack other kinds of NLP tasks than a wordnet, and even work better for some of the same tasks, e.g., *lexical cohesion*, *synonym identification*, *pseudo-word-sense disambiguation*, and *analogy problems* (Morris and Hirst, 1991; Jarmasz and Szpakowicz, 2004; Kennedy and Szpakowicz, 2008; Kennedy and Szpakowicz, 2014).

An obstacle to the wider use of Roget in NLP applications is its limited availability. The only free digital version is the 1911 American edition available through Project Gutenberg.[3] This version is obviously not well suited for processing modern texts. Szpakowicz and his colleagues at the University of Ottawa have conducted a number of experiments with a modern (from 1987) edition of Roget (e.g., Jarmasz and Szpakowicz 2004; Kennedy and Szpakowicz 2008, but as far as we can tell, this dataset is not generally available, due to copyright restrictions. The work reported by Kennedy and Szpakowicz (2014) represents an effort to remedy this situation, utilizing corpus-based measures of semantic relatedness for adding new entries to both the 1911 and 1987 editions of Roget.

In order to investigate systematically the strengths and weaknesses of diverse lexical-semantic resources when applied to different classes of NLP tasks, we would need access to resources that are otherwise comparable, e.g., with respect to language, vocabulary and domain coverage. The

[1] Parts of the introduction reproduced from Borin et al. (2015).

[2] See the *Global WordNet Association* website: <http://globalwordnet.org>.

[3] See <http://www.gutenberg.org/ebooks/22> and Cassidy (2000).

resources should also ideally be freely available, in order to ensure reproducibility as well as to stimulate their widest possible application to a broad range of NLP problems. Unfortunately, this situation is rarely encountered in practice; for English, the experiments contrasting WordNet and Roget have indicated that these resources are indeed complementary. It would be desirable to replicate these findings for other languages and also using lexical-semantic resources with different structures (WordNet and Roget being two out of a large number of possibilities).

This is a central motivation for the work presented here, the ultimate goal of which is to develop automatic methods for producing or considerably facilitating the production of a Swedish counterpart of Roget with a large and up-to-date vocabulary coverage. This is not to be done by translation, as in previous work by de Melo and Weikum (2008) and Borin et al. (2014). Instead, an existing but largely outdated Roget-style thesaurus will provide the scaffolding, where new word senses can be inserted, drawing on the formal structure of an existing Swedish semantic lexicon, Saldo (Borin et al., 2013). Saldo was originally conceived as an "associative thesaurus" (Lönngren, 1998), and even though its organization in many respects differs significantly from that of Roget, there are also some commonalities. Hence, our hypothesis is that the structure of Saldo will yield a good measure for the semantic relatedness of word senses. Saldo is described in Section 2.2 below.

2. The Datasets

2.1. Bring's Swedish Thesaurus

Sven Casper Bring (1842–1931) was the originator of the first and so far only adaptation of Roget's *Thesaurus* to Swedish, which appeared in 1930 under the title *Svenskt ordförråd ordnat i begreppsklasser* 'Swedish vocabulary arranged in conceptual classes' (referred to as "Bring" or "Bring's thesaurus" below). The work itself consists of two parts: (1) a conceptually organized list of Roget categories; and (2) an alphabetically ordered lemma index.

Like in Roget, the vocabulary included in Bring is divided into slightly over 1,000 "conceptual classes". A "conceptual class" corresponds to what is usually referred to as a "head" in the literature on Roget. Each conceptual class consists of a list of words (lemmas), subdivided first into nouns, verbs and others (mainly adjectives, adverbs and phrases), and finally into groups. In the groups, the distance – expressed as difference in list position – between words provides a rough measure of their semantic distance.

Bring thus forms a hierarchical structure with four levels:

(1) conceptual class (Roget "head")
(2) part of speech
(3) group
(4) lemma (word sense)

Since most of the Bring classes have corresponding heads in Roget, it should be straightforward to add the levels above Roget heads/Bring classes to Bring if needed. There are some indications in the literature that this additional structure can in fact be useful for calculating semantic similarity (Jarmasz and Szpakowicz, 2004).

Bring's thesaurus is made available in two digital versions by Språkbanken Text (the text division of the National Swedish Language Bank) at the University of Gothenburg, both versions under a Creative Commons Attribution License:

Bring (v. 1): A digital version of the full contents of the original 1930 book version (148,846 entries).[4]

Blingbring (v. 0.2), a version of Bring where obsolete items have been removed and the remaining entries have been provided with word sense identifiers from Saldo (see section 2.2), providing links to most of Språkbanken Text's other lexical resources. This version contains 126,911 entries.[5]

The linking to Saldo senses in the current Blingbring version (v 0.2) has not involved a disambiguation step. Rather, it has been made by matching lemma-POS combinations from the two resources. For this reason, Blingbring includes slightly over 21,000 ambiguous entries, or about 4,800 ambiguous word sense assignments (out of about 43,000 unique lemma-POS combinations).

The aim of the experiments described below has been to assess the feasibility of disambiguating these ambiguous linkages automatically, and specifically also to evaluate Saldo as a possible knowledge source for accomplishing this disambiguation. The longer-term goal of this work is to develop good methods for adding modern vocabulary automatically to Bring from, e.g., Saldo, thereby hopefully producing a modern Swedish Roget-style resource for the NLP community.

2.2. Saldo

Saldo (Borin et al., 2013) is a large (137 thousand entries and 2 million word forms) morphological and lexical-semantic lexicon for modern Swedish, freely available (under a Creative Commons Attribution license).[6]

As a lexical-semantic resource, Saldo is organized very differently from a wordnet (Borin and Forsberg, 2009). As mentioned above, it was initially conceived as an "associative thesaurus". Since it has been extended following the principles laid down initially by Lönngren (1998), this characterization should still be valid, even though it has grown tremendously over the last decade.

If the fundamental organizing principle of PWN is the idea of full synonyms in a taxonomic concept hierarchy, the basic linguistic idea underlying Saldo is instead that, semantically speaking, the whole vocabulary of a language can be described as having a center – or core – and (consequently) a periphery. The notion of *core vocabulary* is familiar from several linguistic subdisciplines (Borin, 2012). In Saldo this idea is consistently applied down to the level of individual word senses.

The basic lexical-semantic organizational principle of Saldo is hierarchical. Every entry in Saldo – representing a word sense – is supplied with one or more semantic descriptors, which are themselves also entries in the dictionary. All entries in Saldo are actually occurring words or

[4] <https://spraakbanken.gu.se/eng/resource/bring>

[5] <https://spraakbanken.gu.se/eng/resource/blingbring>

[6] <https://spraakbanken.gu.se/eng/resource/Saldo>

conventionalized or lexicalized multi-word units of the language. No attempt is made to fill perceived gaps in the lexical network using definition-like paraphrases, as is sometimes done in PWN (Fellbaum, 1998a, 5f). A further difference as compared to PWN (and Roget-style thesauruses) is that Saldo aims to provide a lexical-semantic description of *all* the words of the language, including the closed-class items (prepositions, conjunctions, interjections, etc.), and also including many proper nouns.

One of the semantic descriptors in Saldo, called *primary*, is obligatory. The primary descriptor is the entry which better than any other entry fulfills two requirements: (1) it is a semantic neighbor of the entry to be described and (2) it is more central than it. However, there is no requirement that the primary descriptor is of the same part of speech as the entry itself. Thus, the primary descriptor of *kniv* 'knife (n)' is *skära* 'cut (v)', and that of *lager* 'layer (n)' is *på* 'on (p)'. Through the primary descriptors Saldo is a single tree, rooted by assigning an artifical top sense (called PRIM) as primary descriptor to the 41 topmost word senses.

That two words are semantic neighbors means that there is a direct semantic relationship between them (such as synonymy, hyponymy, meronymy, argument-predicate relationship, etc.). As could be seen from the examples given above, Saldo includes not only open-class words, but also pronouns, prepositions, conjunctions etc. In such cases closeness must sometimes be determined with respect to function or syntagmatic connections, rather than ("word-semantic") content.

Centrality is determined by means of several criteria: frequency, stylistic value, word formation, and traditional lexical-semantic relations all combine to determine which of two semantically neighboring words is to be considered more central.

For more details of the organization of Saldo and the linguistic motivation underlying it, see Borin et al. (2013).

Like Roget, Saldo has a kind of topical structure, which – again like Roget, but different from a wordnet – includes and connects lexical items of different parts of speech, but its topology is characterized by a much deeper hierarchy than that found in Roget. There are no direct correspondences in Saldo to the lexical-semantic relations making up a wordnet (minimally synonymy and – part-of-speech internal – hyponymy).

Given the (claimed) thesaural character of Saldo, we would expect a Saldo-based semantic similarity measure to work well for disambiguating the ambiguous Blingbring entries.

3. The Experiments

The experiments described below represent a continuation of an earlier effort, reported on by Borin et al. (2015), where both a corpus-based and a lexicon-based classifier was applied to the disambiguation problem, reaching accuracies of 69% and 78%, respectively. The lexicon-based representations used in the earlier experiment utilized only one of several possible aspects of the lexical structure of Saldo, and in the experiments reported here we conduct a more detailed investigation of if and how more of Saldo's structure could be used for this purpose. While these earlier experiments use machine learning, that is, statistical methods, the approach we use here is much simpler and arguably non-statistical. As we will see, it is sometimes possible to get better results with methods simpler than the conventional. There is still a possibility of combining this type of method with a machine learning approach, either in parallel or sequentially, but we leave this for future work.

The evaluation data used for the experiments are the same as in Borin et al. (2015), and we reproduce the data preparation procedure from that paper here for convenience.

The Blingbring data were downloaded from Språkbanken Text's website and a sample of ambiguous Bring–Saldo linkages was selected for manual disambiguation.

An initial sample was drawn from this data set according to the following principles:[7]

- The sampling unit was the class+part of speech-combination, i.e., *nouns in class 12*, *verbs in class 784*, etc.
- This unit had to contain at least 100 lemmas (actual range: 100–569 lemmas),
- out of which at least 1 must be unambiguous (actual range: 56–478 unambiguous lemmas),
- and at least 4 had to be ambiguous.
- From the ambiguous lemmas, 4 were randomly selected (using the Python function random-sample).

The goal was to produce an evaluation set of approximately 1,000 items, and this procedure yielded 1,008 entries to be disambiguated. The disambiguation was carried out by one of the authors. In practice, it deviated from the initial procedure and proceeded more opportunistically, since reference often had to be made to the main dataset in order to determine the correct Saldo word sense. On these occasions, it was often convenient to (a) either disambiguate additional items in the same Bring class; and/or (b) disambiguate the same items throughout the entire dataset.

1,368 entries were disambiguated for the experiments, out of which about 500 came out of the original sample.

For this experiment, a few of those were removed for various anomalies, most commonly because the Bring words are inflected forms and so not directly listed as lemmas in Saldo. This leaves 1317 entries. The degree of ambiguity in this gold standard data is shown in the second column of Table 1, while the third column shows the degree of ambiguity in the full Blingbring dataset containing 44,615 unique lemma-POS combinations.

4. Method and Results

There are two tasks we would like to accomplish. First, there are a number of entries in Bring which are ambiguous, in that they are not associated with one specific Saldo sense. We want to figure out for each of them which of the possible senses is the correct one. Second, there are many entries in Saldo which are not represented in Bring, which we would like to add, so we need to find for each of the Saldo senses which (one or more) of the Bring categories they fit in.

[7] These should be seen as first-approximation heuristic principles, and not based on any more detailed analysis of the data. We expect that further experiments will provide better data on which to base such decisions.

# senses/ entry	GS data: # entries	Blingbring: # entries
1	9	39,275
2	739	4,006
3	304	873
4	147	286
5	71	102
6	11	31
7	13	18
8	15	10
9	6	3
10	2	6
11	0	5

Table 1: Word-sense ambiguity in the gold standard data and in Blingbring

For the first task, we can easily look up which senses in Saldo are associated with the lemmas used in Bring, which already narrows it down to a usually quite small number of possible senses. Most Bring entries have only one possible sense; those are of course not ambiguous and therefore not included in this task. Of the ambiguous ones, most have only two possible senses.

The second task is more difficult. Rather than just a small number of options, we now need to distinguish between several thousand categories. The same sense can also be present in more than one category. In principle, entries in Bring are also ordered in such a way that more similar words are generally closer together. This is difficult to quantify, so we will neither make use of it nor consider it for output.

4.1. Method

Both Bring and Saldo have connections between entries. In Bring, they are arranged in classes and groups; in Saldo, they have primary and secondary descriptors. To predict whether a sense is a good fit for a Bring group, we compare the established entries in the same group with Saldo entries related to the sense at hand.

To compare the different types of relationships between senses in Saldo, we can borrow terminology from family relations. We let the primary descriptor of a sense be its ”mother”, a secondary descriptor its ”father”. A sense which has this one as its primary or secondary is its ”daughter” or ”son”, respectively. Senses sharing a primary or secondary descriptor are ”sisters” or ”brothers”, respectively. In the otherwise rare case where the mother of one sense is the father of another, we will call them ”cross siblings”. Terms like parent, aunt, etc. should follow by analogy.

Many of the Saldo senses have no secondary descriptors, and are therefore ignored when considering ”brothers” etc. We also ignore any secondary descriptor which is inte..1 ‘not’; this links a lot of words which are negations but otherwise have nothing in common.

4.2. Disambiguating Senses of Entries Already Present in Bring

4.2.1. Method

We start with the list of 1317 manually disambiguated Bring entries, as described in Section 3, and find all the Saldo senses which correspond to the same lemma. Both Bring and Saldo give us information on part of speech, although in different forms. In principle, the correct sense could have been listed as having a different part of speech, but we find that this is never the case; consequentially, we remove as candidates all the senses where the part of speech is not the same as that stated in Bring.

The average number of remaining senses is 2.8, and the maximum is 10. This means that if we were to guess a sense at random, we would get an accuracy of 36%. But although the senses in Saldo are not ordered by any formal criterion, they have a tendency to be listed with the more common first. If we choose the first listed sense, we actually get 63% correct. We consider that to be our baseline for accuracy.

Now we process for each of the ambiguous entries each of the possible senses, by considering related senses and seeing if they are present in the same Bring category. To do that, we have to choose on the one hand which type of relations we are considering, and on the other hand which of the two Bring categories to count – classes (the larger) or groups (the smaller).

It quickly becomes clear that some of the relations are stronger indicators than others. For example, if a descriptor of the sense in question is present in the group, that is a very strong indicator, but on the other hand, it only happens in a small percentage of the cases. Conversely, a sense with a shared descriptor appearing in the class is much more common, but is a less strong indicator that this is the correct sense.

This gives us an advantage over a simple discrimination method: We can decide not to make a choice on some cases. If we can get a very high accuracy on, for example, half the entries, that may be much better than just getting a 50% accuracy on all the entries.

It seems therefore like a sensible approach to start with the most accurate but least thorough method, and then apply different methods in turn. That is, if the first method finds a match, that will be our guess, otherwise we move on to the next. If there are several matches, the algorithm stops at the first match, meaning that we get the first listed of the alternatives. If none of the methods work, we also revert to picking the first listed sense.

4.2.2. Results and Discussion

Table 2 and Figure 1 show the results. We can either spot a small number of entries with high accuracy, or a larger number of entries with lower accuracy.

One example of an ambiguous word is *mask*, which shows up in several different groups in Bring. The word has at least two unrelated senses, both nouns: mask..1 translates as ‘worm’, mask..2 as ‘mask’. In our test set, there are three occurrences of what should be mask..2, in the classes AMUSEMENT, DEFENSE, and COVERING. The first is correctly identified because of a son sense; maskerad..1 ‘masquerade’ is in the same group, and has mask..2 as its

Relation	This step %		So far %	
	Tried	Acc	Tried	Acc
father in group	1	100	1	100
mother in group	14	94	14	94
daughter in group	19	90	31	92
son in group	3	80	33	91
grandparent in group	5	85	37	91
sister in group	23	91	51	91
cross sibling in group	9	66	56	89
brother in group	3	60	58	88
sister in class	39	76	74	85
cross sibling in class	29	64	81	83
brother in class	2	80	82	83
father in class	0	100	82	83
mother in class	14	61	84	83
grandparent in class	11	73	86	83
daughter in class	10	72	87	82
son in class	2	75	88	82
first listed option	100	59	100	80

Table 2: Methods for disambiguating Bring entries, and their accuracies, sequentially applied

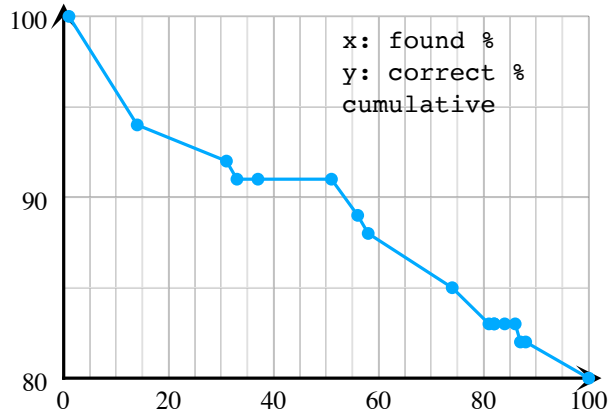

Figure 1: Coverage and accuracy for different methods of disambiguation

secondary descriptor. The second is correctly identified because of a sister sense; överdrag..1 'textile cover' is in the same group, and they share the primary descriptor täcka..1 'cover'. The third is wrongly identified as mask..1 'worm', because of a cross sibling sense; päls..1 'fur' is in the group, and djur..1 'animal' is both the primary descriptor of mask..1 and a secondary descriptor of päls..1.

Generally, most of the failed words, and indeed most of the words altogether, are more closely related senses than this – sometimes clearly distinct but etymologically related senses, including metaphors, such as tomhänt..1 'with empty hands' and tomhänt..2 'with nothing to offer', sometimes with only subtle differences, such as samling..1 'collection', samling..2 'arrangement', and samling..3 'group'.

One obvious alternative approach is to give points for each relative spotted, and check which sense gets the most points. A simple test of this shows no noticeable improvement; further comparison has to be left for future work.

There are other potential extentions to this methods that we could have tried: Reordering the relations, trying additional relations, considering the distance between entries in Bring, considering how far from the root node an entry is in Saldo, looking for combinations of multiple relations occurring in the same category... But preliminary tests show no indication that the real accuracy would be affected by more than a minute amount, and so we leave out further micromanagement to avoid overfitting.

Another possible addition worth considering would be to check the actual frequencies of the senses, and use those instead of the order in Saldo to make the default choice. But without a very large amount of text data, we would not want to rely on the assumption that not only most words but most senses in the dictionary are accurately represented. Manually sense-disambiguated data is somewhat scarce, and we would also not want to rely on automatically sense-disambiguated data; unlike many other applications, we are not interested in the per-token accuracy, but rather the per-lemma accuracy, which is clearly lower, since the sense disambiguation will also be less accurate for less common words.

Relation	Count in sample		Avg. per group		
	True	False	True	False	Ratio
mother	468	3376	0.0969	0.0004	221.2
father	58	694	0.0120	0.0001	133.4
sister	1688	20134	0.3494	0.0026	133.8
brother	635	10465	0.1314	0.0014	96.8
cross sibling	527	10097	0.1091	0.0013	83.3
daughter	701	3156	0.1451	0.0004	354.4
son	130	1651	0.0269	0.0002	125.7
grand-parent	151	5270	0.0313	0.0007	45.7
aunt/ uncle	2313	65570	0.4788	0.0085	56.3
cousin	6753	418069	1.3978	0.0542	25.8

Table 3: Number of occurrences of different relations, for a sample of 10,000 entries

4.3. Adding New Senses to Bring

Now we turn to the second task, in which we want to take senses which are not present in Bring and add them in the correct group. We use the same principles here, looking for groups containing Saldo-relatives of the sense in question.

Is it reasonable to think that a sense will have more relatives in the correct category than in other categories? We test this by counting some types of relatives in different categories. For 10,000 unambiguous entries in Bring, we count the relatives in true groups (that is, any group containing an entry using the same sense), compared with those in false groups (groups which do not contain such an entry). Table 3 shows the results.

We see that there are indeed considerably more relatives in the correct groups. For example, a group that contains a

given sense x will on average contain 0.13 of its brothers, but a group that does not contain x contains only 0.0014 of its brothers.

Does this mean that we can apply the same method as before, and classify any group containing close relatives of x as likely true groups for x? Unfortunately not, since in this task we have far more options to choose from. Of the sense/group combinations in this sample, there are approximately 1600 times as many false ones. So while the mother sense is about 200 times more likely to be found in a true group than a false group, a group containing the mother sense is still 8 times more likely to be a false group.

Instead, we revisit the idea of a scoring system, counting multiple relatives in the same group. This did not seem to improve the sense disambiguation task noticeably, but it might work better here. As we see in Table 3, the more distant relatives have generally less impressive numbers, and preliminary testing also shows that they do not significatly improve results. We limit the method to parent, child and sibling senses, and give one point for each relative.

For each of the Saldo senses associated with an unambiguous Bring entry, we compare it with each of the 7714 Bring groups. For each sense/group combination, we note the score, and whether the group contains the sense itself or not. This tells us the distribution of scores, that is, how many sense/group combinations were given each score.

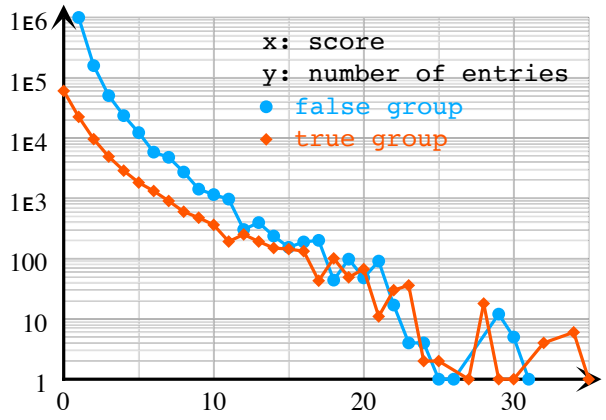

Figure 2: Distribution of scores for true and false groups

4.3.1. Results and Discussion

We find that 24% of the entries were "correctly" classified, that is, the highest-scoring group was a true group. Note that this includes entries which did not get any points in any groups. This in itself is hardly enough accuracy to be useful.

Figure 2 shows the distribution of scores, separately for true and false groups. (Note that one point is outside the graph; there were 301E6 false groups with score 0.) Our hope was that for high enough scores, the true groups would outnumber the false, so that beyond a certain score limit we might have a decent accuracy. As we see in the graph, the false groups remain higher at least up to score 10; after that, the smaller number of data points make the graph more erratic.

Figure 3 shows the percentage of true groups for each score. The blue curve shows the percentage of true groups among

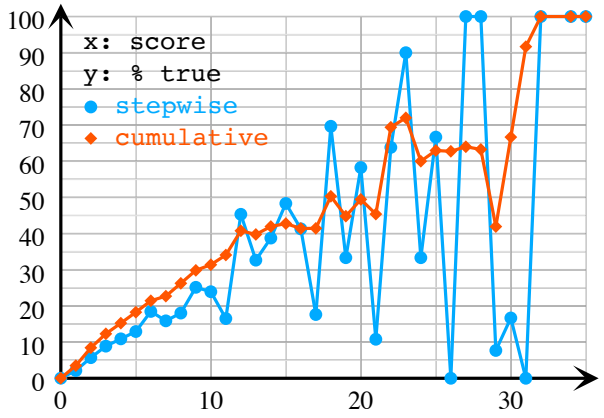

Figure 3: Percentage of true groups for each score. The blue line is for exactly this score, the orange is for at least this score

those with exactly this score, and the orange curve shows the percentage among groups with this score or higher. We see that the percentage does increase noticeably in the lower part. Already beyond 10 or so, the results are less reliable, but the general trend seems to be increasing.

If we were to set a score limit and assign senses to groups if they reach that limit, the orange curve would describe the accuracy of that method. As far as we can tell, this would reach an accuracy around 30% at 10 points. Unfortunately, this method would not be satisfactory. First, an accuracy of 30% is not good enough. Second, the method would only attempt a very small number of words; only one in 200,000 sense/group combinations score at least 10 points.

On average, each word in Bring appears in 2.88 categories, but we would be satisfied for now with finding just one for each new word. Since the automatic methods are not accurate enough, we need to try semi-automatic methods. What if we set a lower score limit, and manually go though the categories with a sufficient score? If we could narrow it down to a list of ten or even a hundred candidate groups instead of the full list of 7714, that would be very helpful.

With a score limit of just 1, the accuracy is 3.5%, and the recall is 43.6% (that is, out of all the true groups, we will find 43.6% by looking at those with at least 1 point). With a score limit of 2, the accuracy is 8.4% and the recall 22.7%. This may be better than nothing, but still not overwhelming.

Instead, we can choose to list the suggested groups in order of decreasing score, and see how many groups we would on average need to look at to find a true group. Figure 4 shows the result.

We see that while 24% are found in the first guess, 43% are found in the first 5, and 50% in the first 10. That should at least be enough to reduce the workload of an annotator. Even if the first few listed groups are not correct, it might also give the annotator an idea of where to look – other groups in the same class would presumably be more likely than more distant ones.

5. Conclusions

We have shown that using the relations from Saldo to disambiguate or classify words in Bring is viable as a tool,

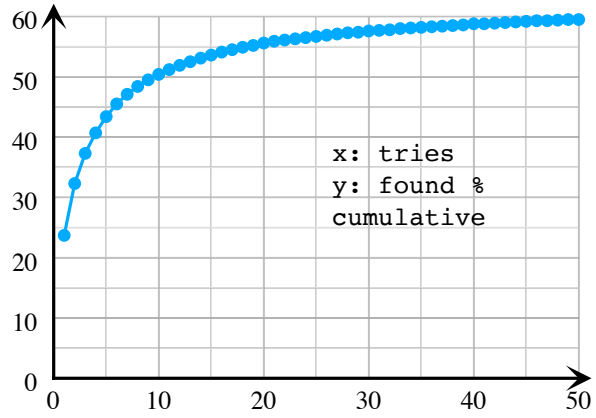

Figure 4: Percentage of entries for which a true group is found within a given number of groups, starting from the highest-scoring

even if the accuracy is not high enough to rely solely on this method. For disambiguation of already existing entries, we can get an accuracy or 80% for the entire list, and higher for a subset; this may be considered acceptable in itself, or it can be seen as a starting point for manual annotators. For classifying new senses, the accuracy is not good enough for automatic annotation, but it can reduce the number of possible groups a manual annotator would have to look through by a factor of several hundred.

It is important to note that the correct answer here is somewhat subjective. There may be cases where a different sense would be just as reasonable, and perhaps more importantly, there are many cases where more than one sense would fit in the same category. Some of the words in Bring are clear homographs, so the senses are very different and should clearly be in different categories, but others may be more closely related senses. This means that the accuracies we see here might be overly pessimistic.

Given more time and resources, it would be possible to extend the manual annotation which we have used as our gold standard. Having more than one annotator might give us a better picture of just how subjective the annotation is, and an approach where for each included sense we also classify the other senses of the same word would perhaps clarify whether the accuracy is actually better than it seems.

It is also possible to combine the approach presented here with other automatic methods, whether commonplace machine learning methods or something else, which is something we intend to do in the future. All the same, we have shown that these transparent, conceptually simple, and relatively fast methods are also quite viable.

6. Acknowledgements

This work has been conducted as part of the effort to construct and develop a Swedish national research infrastructure in support of research based on language data. This infrastructure – *Nationella språkbanken* (the Swedish National Language Bank) – is jointly funded for the period 2018–2024 by the Swedish Research Council (grant number 2017-00626) and its 10 partner institutions.

7. Bibliographical References

Borin, L. and Forsberg, M. (2009). All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources*, Odense.

Borin, L., Forsberg, M., and Lönngren, L. (2013). SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Borin, L., Allwood, J., and de Melo, G. (2014). Bring vs. MTRoget: Evaluating automatic thesaurus translation. In *Proceedings of LREC 2014*, pages 2115–2121, Reykjavík. ELRA.

Borin, L., Nieto Piña, L., and Johansson, R. (2015). Here be dragons? The perils and promises of inter-resource lexical-semantic mapping. In *Semantic Resources and Semantic Annotation for Natural Language Processing and the Digital Humanities. Workshop at NODALIDA 2015*, pages 1–11, Linköping. LiUEP.

Borin, L. (2012). Core vocabulary: A useful but mystical concept in some kinds of linguistics. In Diana Santos, et al., editors, *Shall we play the Festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday*, pages 53–65. Springer, Berlin.

Cassidy, P. (2000). An investigation of the semantic relations in the Roget's Thesaurus: Preliminary results. In *Proceedings of CICLing 2000*, pages 181–204.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press, Cambridge.

de Melo, G. and Weikum, G. (2008). Mapping Roget's Thesaurus and WordNet to French. In *Proceedings of LREC 2008*, Marrakech. ELRA.

de Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York. ACM.

Fellbaum, C. (1998a). Introduction. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 1–19. MIT Press, Cambridge, Mass.

Christiane Fellbaum, editor. (1998b). *WordNet: An electronic lexical database*. MIT Press, Cambridge, Mass.

Goddard, C. (2001). Lexico-semantic universals: A critical overview. *Linguistic Typology*, 5:1–65.

Hüllen, W. (2004). *A history of Roget's Thesaurus: Origins, development, and design*. Oxford University Press, Oxford.

Jarmasz, M. and Szpakowicz, S. (2004). *Roget's Thesaurus* and semantic similarity. In Nicolas Nicolov, et al., editors, *Recent Advances in Natural Language Processing III. Selected papers from RANLP 2003*, pages 111–120. John Benjamins, Amsterdam.

Jobbins, A. C. and Evett, L. J. (1995). Automatic identification of cohesion in texts: Exploiting the lexical organization of Roget's Thesaurus. In *Proceedings of Rocling VIII*, pages 111–125, Taipei.

Jobbins, A. C. and Evett, L. J. (1998). Text segmentation using reiteration and collocation. In *Proceedings of the 36th ACL and 17th COLING, Volume 1*, pages 614–618, Montreal. ACL.

Kennedy, A. and Szpakowicz, S. (2008). Evaluating *Roget's* thesauri. In *Proceedings of ACL-08: HLT*, pages 416–424, Columbus, Ohio. ACL.

Kennedy, A. and Szpakowicz, S. (2014). Evaluation of automatic updates of *Roget's Thesaurus*. *Journal of Language Modelling*, 2(2):1–49.

Lönngren, L. (1998). A Swedish associative thesaurus. In *Euralex '98 proceedings, Vol. 2*, pages 467–474.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Murphy, M. L. (2003). *Semantic relations and the lexicon*. Cambridge University Press, Cambridge.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roget, M. P. (1852). *Thesaurus of English Words and Phrases*. Longman, London.

Martine Vanhove, editor. (2008). *From polysemy to semantic change: Towards a typology of lexical semantic associations*. Jon Benjamins, Amsterdam.

Wilks, Y. (1998). Language processing and the thesaurus. In *Proceedings National language Research Institute*, Tokyo. Also appeared as Technical report CS–97–13, University of Sheffield, Department of Computer Science.