Here be dragons?
The perils and promises of inter-resource lexical-semantic mapping

Lars Borin **Luis Nieto Piña** **Richard Johansson**

Språkbanken, Department of Swedish, University of Gothenburg, Sweden
{lars.borin, luis.nieto.pina, richard.johansson}@svenska.gu.se

Abstract

Lexical-semantic knowledges 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.

In linguistics, lexical semantics comes in many flavors, but in the NLP world, wordnets reign more or less supreme. There has been some promising work utilizing Roget-style thesauruses instead, but wider experimentation is hampered by the limited availability of such resources.

The work presented here is a first step in the direction of creating a freely available Roget-style lexical resource for modern Swedish. Here, we explore methods for automatic disambiguation of inter-resource mappings with the longer-term goal of utilizing similar techniques for automatic enrichment of lexical-semantic resources.

1 Introduction

1.1 The uniformity of lexical semantic resources for NLP

Lexical-semantic knowledges 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 certainly endeavor to 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.[1]

Manually built alternatives to wordnets are afflicted by being for some other language than English (e.g., SALDO: Borin et al. 2013) or by not being freely available – see the next section – or possibly both.

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

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

Lars Borin, Richard Johansson and Luis Nieto Piña 2015. Here be dragons? The perils and promises of inter-resource lexical-semantic mapping. *Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015.* NEALT Proceedings Series 27 / Linköping Electronic Conference Proceedings 112: 1–11.

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).

It has been proposed in the literature that Roget-style thesauruses could 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.[2] 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 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, e.g., 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 certainly 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 with the help of two different kinds of semantic relatedness measures:

1. One such measure is corpus-based, similar to the experiments conducted by Kennedy and Szpakowicz (2014), described above.
2. The other measure utilizes an existing lexical-semantic resource (SALDO: Borin et al. 2013).

In the latter case, we also have a more theoretical aim with our work. 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 semantic relatedness measure for the task at hand. 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.

In addition, there is a brief preface by S. C. Bring, which we reproduce here in its entirety:[3]

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

[3] This English translation comes from the Bring resource page at Språkbanken: <http://spraakbanken.gu.se/eng/resource/bring>.

> This wordlist has been modelled on P. M. Roget's "Thesaurus of English Words and Phrases". This kind of wordlist can be seen as a synonym dictionary of sorts. But each conceptual class comprises not only synonyms, but words of all kinds which are habitually used in discoursing on the kind of topics which could be subsumed under the class label concept, understood in a wide sense.
>
> Regarding Roget's classification system, there are arguably a number of classes which ought to be merged or split. But this classification seems to have established itself solidly through many editions of Roget's work as well as German copies of it. It should also be considered an advantage that the same classification is used in such dictionaries for different languages.
>
> Uppsala in September 1930.
>
> *S. C. Bring*

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 paragraphs. In the paragraphs, 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) paragraph
(4) lemma (word sense)

This stands in contrast to Roget, where the formal structure defines a nine-level hierarchy (Jarmasz and Szpakowicz, 2001; Jarmasz and Szpakowicz, 2004):

(1) class
(2) section
(3) subsection
(4) category, or head group
(5) head (Bring "conceptual class")
(6) part of speech
(7) paragraph
(8) semicolon group
(9) 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 has recently been made available in two digital versions by Språkbanken (the 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. 0.1), 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's other lexical resources. This version contains 126,911 entries.[5]

The linking to SALDO senses in the current Blingbring version (v 0.1) 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 (out of approximately 127,000 in total), 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 (137K entries and 2M wordforms) 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".

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

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

[6] <http://spraakbanken.gu.se/eng/resource/saldo>

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, as we will now describe.

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 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 thesuruses) is that SALDO aims to provide a lexical-semantic description of *all* the words of the language, including the closed-class items (prepositions, subjunctions, 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 desribed 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, and not be inferior to a corpus-based or wordnet-based measure. There is no sufficiently large Swedish wordnet at present, so for now we must restrict ourselves to a comparison of a corpus-based and a SALDO-based method.

The experiments described below were conducted using SALDO v. 2.3 as available for downloading on Språkbanken's website.

3 Automatic disambiguation of ambiguous Bring entries

We now turn to the question of automatically linking the Bring and SALDO lexicons: many entries in Bring have more than one sense in SALDO, and we present a number of methods to automatically rank SALDO senses by how well they fit into a particular Bring class. Specifically, since entries in Bring are not specified in terms of a sense, this allows us to predict the SALDO sense that is most appropriate for a given Bring entry. For instance, the lexicon lists the noun *broms* as belonging to Bring class 366, which contains a large number of terms related to animals. SALDO defines two senses for this word: *broms-1* 'brake' and *broms-2* ‘horsefly’, but it is only the second sense that should be listed in this Bring class.

In this work we consider the task of selecting a SALDO sense for a Bring entry, but we imagine that the methods proposed here can be applied in other scenarios as well. For instance, it is possible that they could allow us to predict the Bring class for a word that is *not* listed in Bring, but we leave this task for future investigation. The methods are related to those presented by Johansson (2014) for automatically suggesting FrameNet frames for SALDO entries.

We first describe how we use the SALDO network and cooccurrence statistics from corpora to represent the meaning of SALDO entries. These meaning representations are then used to carry out the disambiguation. We investigate two distinct ways to use the representations for disambiguating: (1) by selecting a *prototype* (centroid) for each class, and then selecting the SALDO sense that is most similar to the prototype; (2) by using the existing Bring entries as training instances for a *classifier* that assigns a Bring class to a SALDO entry, and then ranking the SALDO senses by the probability output by the classifier when considering each sense for a Bring class.

3.1 Representing the meaning of a SALDO entry

To be able to connect a SALDO entry to a Bring class, we must represent its *meaning* in some structured way, in order to relate it to other entries with a similar meaning. There are two broad approaches to representing word meaning in NLP work: representations based on the structure of a formal knowledge representation (in our case the SALDO network), and those derived from cooccurrence statistics in corpora (*distributional* representations). In this work, we explore both options.

3.1.1 Word senses in Bring and in SALDO

But even if we restrict ourselves to how they are conceived in the linguistic literature, word senses are finicky creatures. They are obviously language-dependent, strongly so if we are to believe, e.g., Goddard (2001). Furthermore, there seems to be a strong element of tradition – or ideology – informing assumptions about how word senses contribute to the interpretation of complex linguistic items, such as productive derivations, compounds and incorporating constructions, as well as phrases and clauses. This in turn determines the granularity – the degree of polysemy – posited for lexical entries.

One thing that seems to be assumed about Roget – and which if true consequently ought to hold for Bring as well – is that multiple occurrences of the same lemma (with the same part of speech) represent different word senses (e.g., Kwong 1998; Nastase and Szpakowicz 2001). This is consistent with a “splitting” approach to polysemy, similar to that exhibited by PWN and more generally by an Anglo-Saxon lexicographical tradition.

However, this is not borne out by the Bring–SALDO linking. First, there are many unambiguous – in the sense of having been assigned only one SALDO word sense – Bring lemma-POS combinations that appear in multiple Bring classes. Second, during the practical disambiguation work conducted in order to prepare the evaluation dataset for the experiments described below, the typical case was not – as would have been expected if the above assumption were correct – that ambiguous items occurring in several Bring classes would receive different word sense assignments. On the contrary, this turned out to be very much a minor phenomenon.

A “word sense” is not a well-defined notion (Kilgarriff, 1997; Hanks, 2000; Erk, 2010; Hanks, 2013), and it may well be simply that this is what we are seeing here. Specifically, the Swedish lexicographical tradition to which SALDO belongs reflects a “lumping” view on word sense discrimination. If we aspire to link resources such as Roget, Bring, SALDO, etc. between languages, issues such as this need to be resolved one way or another, so there is clearly need for more research here.

3.1.2 Lexicon-based representation

In a structure-based meaning representation, the meaning of a concept is defined by its relative position in the SALDO network. How do we operationalize this position as a practical meaning representation that can be used to compute similarity of meaning or exemplify meaning for a machine learning algorithm? It seems clear that the way this operationalization is carried out has implications for the ability of automatic systems to generalize from the set of SALDO entries associated with a Bring class, in order to reason about new entries.

When using a semantic network, the meaning of a word sense s is defined by how it is related to other word senses; in SALDO, the immediate neighborhood of s consists of a primary descriptor and possibly a set of secondary descriptors, and the meaning of s can be further analyzed by following primary and secondary edges in the SALDO graph. In this work, we follow the approach by Johansson (2014) and let the lexicon-based meaning representation $\phi(s)$ of a SALDO entry s be defined in terms of the transitive closure of the primary descriptor relation. That is, it consists of all SALDO entries observed when traversing the SALDO graph by following primary descriptor edges from s to the SALDO root entry (excluding the root itself). For instance, the meaning of the fourth sense of *fil* 'file (n)' would be represented as the set

$\phi($*fil-4*$)$ = { *fil-4* '(computer) file (n)', *datorminne-1* 'computer memory (n)', *datalagring-1* 'data storage (n)', *lagring-1* 'storage (n)', *lagra-1* 'store (v)', *lager-2* 'stock/store (n)', *förråd-1* 'store (n)', *förvara-1* 'store/keep (v)', *ha-1* 'have (v)' }.

Computationally, these sets are implemented as high-dimensional sparse vectors, which we normalize to unit length. Although in this work we do not explicitly use the notion of similarity functions, we note that the cosine similarity applied to this representation gives rise to a network-based measure similar in spirit to that proposed by Wu and Palmer (1994):

$$\mathrm{sim}(s_1, s_2) = \frac{|\phi(s_1) \cap \phi(s_2)|}{\sqrt{|\phi(s_1)|} \cdot \sqrt{|\phi(s_2)|}}$$

3.1.3 Corpus-based representation

Corpus-based meaning representations rely on the distributional hypothesis, which assumes that words occurring in a similar set of contexts are also similar in meaning (Harris, 1954). This intuition has been realized in a very large number of algorithms and implementations (Turney and Pantel, 2010), and the result of applying such a model is typically that word meaning is modeled *geometrically* by representing co-occurrence statistics in a vector space: this makes it straightforward to define similarity and distance measures using standard vector-space metrics, e.g. the Euclidean distance or the cosine similarity. In this work, we applied the skip-gram model by Mikolov et al. (2013), which considers co-occurrences of each word in the corpus with other words in a small window; this model has proven competitive in many evaluations, including the frame prediction task described by Johansson (2014).

Since our goal is to select a word sense defined by SALDO, but corpus-based meaning representation methods typically do not distinguish between senses, we applied the postprocessing algorithm developed by Johansson and Nieto Piña (2015) to convert vectors produced by the skip-gram model into new vectors representing SALDO senses. For instance, this allows us to say that for the Swedish noun *fil*, the third sense defined in SALDO ('sour milk') is geometrically close to *milk* and *yoghurt* while the fourth sense ('computer file') is close to *program* and *memory*. This algorithm decomposes vector-based word meaning representations into a convex combination of several components, each representing a sense defined by a semantic network such as SALDO. The vector representations of senses are selected so that they minimize the geometric distances to their neighbors in the SALDO graph. The authors showed that the decomposed representations can be used for predicting FrameNet frames for a SALDO sense.

3.2 Disambiguating by comparing to a prototype

The fact that corpus-based representations for SALDO senses are located in a real-valued vector space allows us to generate a prototype for a certain Bring conceptual class by means of averaging the sense vectors belonging to a that class in Bring. This prototype is in the same vector space that the sense representations, so we are able to measure distances between sense vectors and prototypes and determine which sense is closer to the concept embodied in the class prototype.

Thus, our first method for disambiguating links between Bring items and SALDO senses works as follows. For each class j , a prototype c_j is calculated by averaging those sense vectors v_i that are unambiguously linked to a Bring item b_i from class j:

$$c_j = \frac{1}{n} \sum_{b_i \in j} v_i$$

where n is the number of unambiguous links in class j.

Then, for an ambiguous link between a Bring item b_k in class j and its set of possible vectors $\{v_{kl}\}$, the distance from each vector to the class centroid c_j is measured, and the closest one is selected as the representation of the SALDO sense linked to b_k:

$$\arg\min_{l} \mathrm{d}(c_j, v_{kl})$$

where d is a distance function. In our case we have chosen to use *cosine distance*, which is commonly applied on the kind of representations obtained from the skip-gram model (Mikolov et al., 2013) to compute similarity between representations.

3.3 Disambiguating with classifiers

Statistical classifiers offer a wide range of options to learn the distribution of labeled data, which afterwards can be used to label unseen data instances. They are not constrained to work with data in a geometric space, as opposed to the method explained in the previous section. Thus, we can apply classifiers on lexicon-based representations as well.

In our case, we are not interested so much in classifying new instances as in assessing the confidence of such classifications. Consequently, in our ambiguous data we have a set of instances that can possibly be linked to a Bring entry whose class is known to us. Therefore, we would like to ascertain how confident a classifier is when assigning these instances to their corresponding class, and base our decision to disambiguate the link on this information.

For this task we use the Python library Scikit-learn (Pedregosa et al., 2011), a general machine learning package which offers a variety of statistical classifiers. Specifically, we work with a logistic regression method (instantiated with the library's default values, except the inverse regularization strength, set to 100), which classifies instances based on the probability that they belong to each possible class.

The classifier is trained on the set of SALDO sense vectors unambiguously linked to Bring items and their conceptual class information. Once trained, it can be given a set of SALDO sense representations $\{v_{kl}\}$ ambiguously assigned to one Bring entry b_k in class j and, instead of simply classifying them, output their probabilities $\{p_{jl}\}$ of belonging to class j. We then only have to select the sense with the highest probability to disambiguate the link:

$$\arg\max_{l} p_{jl}$$

4 Experiments

4.1 Evaluation data preparation

The Blingbring data was downloaded from Språkbanken'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 the first author. 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.

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

On the other hand, unambiguous entries in Blingbring linking one Bring item to one SALDO sense are isolated to serve as training data. As mentioned above in Section 3.1.1, the structure of Bring's thesaurus makes it possible for a word to appear in more than one conceptual class; if the

[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	–	39,275
2	888	4,006
3	266	873
4	122	286
5	56	102
6	18	31
7	10	18
8	7	10
9	1	3
10	–	6
11	–	5

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

SALDO sense related to those two or more instances is the same, we may have a training instance that spans more than just one class. Initially, it may seem reasonable to exclude such instances from the training data, as their presence may be problematic for the definition of a class. But this phenomenon is quite ubiquitous: 72.6% of the senses unambiguously associated with a Bring entry in Blingbring appear in more than one class. For this reason, we define two different training sets, one that includes *overlap* among the classes and one that does not, and conduct conduct experiments separately on each of them.

4.2 Prototype-based disambiguation

In this section we give the results obtained with the method described in Section 3.2. This experiment is performed using corpus-based representations only, as lexicon-based ones lack a geometrical interpretation, on which the cosine similarity measure used is based.

Table 2 lists the accuracy of the method on our evaluation set. Two results are given corresponding to the training set containing or not instances that span several classes. The accuracy of a random baseline is also given as a reference. Both of the approaches have an accuracy well above the random baseline with an improvement of over 0.14 points, and we observe that there is practically no difference between them, although the approach in which instances overlapping classes are included in the training data performs slightly better.

Method	**Accuracy**
Random baseline	0.4238
Corpus-based, incl. overlap	0.5731
Corpus-based, no overlap	0.5651

Table 2: Disambiguation accuracy using a similarity measure.

PoS	**Proportion**	**Accuracy**
Noun	54.8%	0.5819
Verb	21.3%	0.5538
Others	23.2%	0.5485

Table 3: Disambiguation accuracy by Part-of-Speech using a similarity measure. Overlapping instances included in the training set.

In Table 3 we present for this last case a breakdown of the accuracy into the parts of speech that Bring classes list: *nouns*, *verbs* and *others*.[8] The table also lists the proportions of these classes in the data. No significant difference can be appreciated between the diverse types of words, although nouns fare slightly better than the other two cases.

4.3 Classification-based disambiguation

The results of applying the method introduced in Section 3.3 are given here. In this experiment we also consider lexicon-based data besides the corpus-based representations.

Table 4 lists the accuracies obtained in each instance: corpus-based or lexicon-based data, using either overlapping instances or not. The random baseline accuracy is also shown for reference.

In this case, we observe a greater improvement over the baseline than in the previous experiment with an increase in accuracy of 0.23 between the best cases in each experiment. There is also a considerable difference between the two types of data: the best case using lexicon-based representations provides an accuracy improvement of 0.12 over the best result obtained with corpus-based data. Contrary to the experience of the previous experiment, there is a substantial difference between the presence or absence of overlapping instances in the training data: the accuracy increases by 0.03 in the case of corpus-based data when overlapping instances are used, and by 0.13 in the case of lexicon-based data. This behaviour may seem

[8] As explained in Section 2.1, the tag *others* encompasses mainly adjectives, adverbs and phrases, and unfortunately there is not enough information in Bring to separate these classes and give a more fine-grained analysis.

Method	Accuracy
Random baseline	0.4238
Corpus-based, incl. overlap	0.6879
Corpus-based, no overlap	0.6572
Lexicon-based, incl. overlap	0.7836
Lexicon-based, no overlap	0.6499

Table 4: Disambiguation accuracy using a classifier.

PoS	Accuracy
Corpus-based representations	
Noun	0.7372
Verb	0.6308
Others	0.5825
Lexicon-based representations	
Noun	0.7885
Verb	0.8154
Others	0.7282

Table 5: Disambiguation accuracy by Part-of-Speech using a classifier. Overlapping instances included in the training data.

counter-intuitive, since using training instances that belong to more than one class should dilute the boundaries between those classes. It should be noted here, however, that, given a new instance, the main task assigned in our problem to the classifier is not to decide to which class the instance belongs (as this information is already known), but to output the membership probability for a certain class, so that we are able to compare with those of other instances. Thus, the boundaries between classes matter less to us than the amount of training data that allows the classifier to learn the definition of each class separately.

Table 5 presents an accuracy breakdown for the highest scoring approach in the previous results (i.e., including overlap) using each type of data. These results also differ from the ones in the previous experiments, as we observe a marked difference between parts of speech: using corpus-based representations, nouns obtain the highest accuracy with 0.10 points over the other two classes, while using lexicon based data favours verbs, although closely followed by nouns.

5 Conclusions and future work

Summing up the main results, (1) both the corpus-based and the lexicon-based methods resulted in a significantly higher disambiguation accuracy compared to the random baseline; (2) contrary to intuition, using overlapping instances yielded better accuracy than using only non-overlapping items, which we attribute to the increased amount of training data in the former case; and (3) the hypothesis that the SALDO-based method would yield a better result was supported by the experiments.

The results of the lexicon-based method are already good enough overall that it will be possible to use it as a preprocessing step in order to speed up the disambiguation of the remaining ambiguous entries considerably. The results could also be analyzed in more detail in order to find out whether there are special cases that could be automatically identified where the accuracy may be even higher.

For instance, it would be useful to see whether the structure of the thesaurus can be used in a more sophisticated way. In this work we have only considered the top-level Bring class when selecting among the alternative SALDO senses for an ambiguous Bring entry, but as described in Section 2.1, the thesaurus is organized hierarchically, and closely related terms are placed near each other on the page.

In future work, we would like to investigate to what extent the methods that we have proposed here can be generalized to other Bring-related tasks. In particular, it would be useful to propose a Bring class for words in SALDO that are not listed in Bring, for instance because the word did not exist when the Bring lexicon was compiled. This would make a new and very useful lexical-semantic resource available for use in sophisticated Swedish NLP applications.

Acknowledgements

This research was funded by the Swedish Research Council under the grants *Towards a knowledge-based culturomics* (dnr 2012-5738), *Distributional methods to represent the meaning of frames and constructions* (dnr 2013-4944), and *Swedish FrameNet++* (dnr 2010-6013). We also acknowledge the University of Gothenburg for its support of the Centre for Language Technology and Språkbanken.

References

Lars Borin and Markus Forsberg. 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.

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

Lars Borin, Jens Allwood, and Gerard de Melo. 2014. Bring vs. MTRoget: Evaluating automatic thesaurus translation. In *Proceedings of LREC 2014*, pages 2115–2121, Reykjavík. ELRA.

Lars Borin. 2012. Core vocabulary: A useful but mystical concept in some kinds of linguistics. In Diana Santos, Krister Lindén, and Wanjiku Ng'ang'a, editors, *Shall we play the Festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday*, pages 53–65. Springer, Berlin.

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

D. Alan Cruse. 1986. *Lexical semantics*. Cambridge University Press, Cambridge.

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

Gerard de Melo and Gerhard Weikum. 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.

Katrin Erk. 2010. What is word meaning, really? (And how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 17–26, Uppsala. ACL.

Christiane Fellbaum. 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.

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

Patrick Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1–2):205–215.

Patrick Hanks. 2013. *Lexical analysis. Norms and exploitations*. MIT Press, Cambridge, Massachusetts.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23).

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

Mario Jarmasz and Stan Szpakowicz. 2001. The design and implementation of an electronic lexical knowledge base. In *Proceedings the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI 2001*, pages 325–333.

Mario Jarmasz and Stan Szpakowicz. 2004. *Roget's Thesaurus* and semantic similarity. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing III. Selected papers from RANLP 2003*, pages 111–120. John Benjamins, Amsterdam.

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

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

Richard Johansson and Luis Nieto Piña. 2015. Embedding a semantic network in a word space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, Denver, United States. To appear.

Richard Johansson. 2014. Automatic expansion of the Swedish FrameNet lexicon. *Constructions and Frames*, 6(1):92–113.

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

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

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Oi Yee Kwong. 1998. Aligning WordNet with additional lexical resources. In *Workshop on usage of WordNet in natural language processing systems at COLING-ACL'98*, pages 73–79, Montréal. ACL.

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

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations, Workshop Track*, Scottsdale, USA.

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

M. Lynne Murphy. 2003. *Semantic relations and the lexicon*. Cambridge University Press, Cambridge.

Vivi Nastase and Stan Szpakowicz. 2001. Word-sense disambiguation in Roget's Thesaurus using WordNet. In *Workshop on WordNet and other lexical resources at NAACL*, Pittsburgh. ACL.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mark Peter Roget. 1852. *Thesaurus of English Words and Phrases*. Longman, London.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

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

Yorick Wilks. 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.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA.