Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 51–60
Marseille, 20-25 June 2022

Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR Lexicon

Bolette S. Pedersen[1], Nathalie Carmen Hau Sørensen[1], Sanni Nimb[2], Ida Flörke[2], Sussi Olsen[1], Thomas Troelsgård[2]

Centre for Language Technology, NorS, University of Copenhagen[1], Society for Danish Language and Literature [2]
Emil Holms Kanal 2, 2300 Copenhagen S[1], Christian Brygge 1, 1219 Copenhagen K [2]
{bspedersen, nmp828, saolsen}@hum.ku.dk, {sn,if,tt}@dsl.dk

Abstract

We present the Central Word Register for Danish (COR), which is an open source lexicon project for general AI purposes funded and initiated by the Danish Agency for Digitisation as part of an AI initiative embarked by the Danish Government in 2020. We focus here on the *lexical semantic* part of the project (COR-S) and describe how we – based on the existing fine-grained sense inventory from Den Danske Ordbog (DDO) – compile a more AI suitable sense granularity level of the vocabulary. A three-step methodology is applied: We establish a set of linguistic principles for defining *core senses* in COR-S and from there, we generate a hand-crafted gold standard of 6,000 lemmas depicting how to come from the fine-grained DDO sense to the COR inventory. Finally, we experiment with a number of language models in order to automatize the sense reduction of the rest of the lexicon. The models comprise a ruled-based model that applies our linguistic principles in terms of features, a word2vec model using cosine similarity to measure the sense proximity, and finally a deep neural BERT model fine-tuned on our annotations. The rule-based approach shows best results, in particular on adjectives, however, when focusing on the average polysemous vocabulary, the BERT model shows promising results too.

Keywords: computational lexicon, sense granularity, semantic clustering, semantic text similarity

1. Introducing COR – an open source lexicon for AI purposes

Companies and public institutions in Denmark are right now entering the field of language-centered AI – and are therefore working intensively with Danish language data from an NLP perspective. This development has led to an increased request for a standard machine-readable lexicon of Danish with first of all basic morphology and semantics (core senses, sentiment etc.)[1]. Even if several computational lexical resources for Danish already exist (cf. the language technology portal sprogteknologi.dk under the Danish Agency for Digitization), their license properties as well as their coverage and quality vary quite a lot.

New players in the field also indicate that it is time-consuming for the developers to get an overview of existing resources and their availability and specifically to get an overview of coverage, validity, and actual usefulness of the resource for a particular task. Further, coordination with a company's existing terminology is often complex. These difficulties seem to be enforced by the fact that there is today often a lack of in-house linguistic expertise in companies and public institutions who can guide the use and adjustment of such resources.

To meet these needs, Det Centrale OrdRegister (The Central Word Register of Danish, COR) was initiated in 2021 as part of a Danish governmental LT and AI initiative. The overall aim of the project is to help boost NLP and language-centric AI for Danish. It is funded by The Danish Agency for Digitisation and led in collaboration by three of the main dictionary and LT institutions in Denmark: The Danish Language Council (DSN), Society for Danish Language and Literature (DSL), and The Centre for Language Technology (CST) at the University of Copenhagen.

In COR we aim at compiling a coordinated and standardized framework for machine readable lexical resources for Danish where all lemmas are assigned *a unique identifier*.[2][3] The idea is that company specific terminologies can be added subsequently and given their own unique series of identifiers. Either as open source like the main resource, or with a more restricted license.

The main COR resource consists of a lexicon of the general language vocabulary with basic morphology and semantics. Syntax and phonology are foreseen in a subsequent second phase.

Another main idea is that COR is based on *high-quality, locally anchored knowledge about the Danish language and society* as typically depicted in national dictionaries and thesauri, and not transferred and subsequently adjusted from similar English AI and NLP resources, as it is sometimes seen. Intellectual Property Rights issues are often preventing the inclusion of lexicographic data into open access resources, but in the ELEXIS project these issues have been addressed (Kosem et al., 2021). Based on knowledge on the practices among the other partners in ELEXIS, DSL made clear decisions about which type of information in the national dictionaries compiled by the Society would be allowed in an open-source lexicon like COR.

[1] See Kirchmeier et al. (2019) for a status on Danish language technology resources where this request is put forward.

[2] Which can be seen as a parallel to The Danish Person Register (CPR) where all Danish citizens are assigned a unique id.

[3] See also the COR description on website of The Danish Language Council: https://dsn.dk/nyheder-og-arrangementer/dansk-sprognaevn-med-i-stor-sprogteknologisk-satsning/

Four background resources, which are linked to one another at DDO sense level, constitute the skeleton of the semantic part of COR (COR-S):

- The Danish Dictionary (DDO)
- The Danish wordnet, DanNet
- The Danish Thesaurus (DT), and
- The Danish FrameNet

The Danish Dictionary describes more than 100,000 lemmas of modern Danish, illustrated with corpus examples, collocations, synonyms, usage information etc. The data is organized in a well-structured XML format with unique identifiers on all sense definitions. The dictionary has been published online since 2009 and is continuously being extended with new lemmas and senses.

DanNet (Pedersen et al., 2009) was built on the DDO meaning that, instead of – as most wordnets do – compiling the wordnet as a transfer and adjustment of Princeton WordNet, it is based on monolingual grounds and subsequently linked to Princeton WordNet (see Pedersen et al., 2019). The sense definitions from the DDO were semi-automatically transformed into wordnet relations. The rather fine-grained sense inventory of DDO was more or less taken over in DanNet, however in a 'classical' wordnet manner, that is, with unstructured senses and thus not capturing the structure of main and sub-senses from the DDO and not necessarily all its senses. In cases of synonymy, a wordnet approach was adopted of typically including synonyms as part of the same synset.

The Danish Thesaurus is also based on the DDO and presents approx. 95% of its lemmas and senses in at least one of 888 named thematic sections, listed together with synonyms and near-synonyms in a semantic group initiated by a keyword (henceforth DT keyword). In the compilation process of the thesaurus, the DanNet hierarchies and ontological types supplemented the information extracted from DDO.

Finally, The Danish FrameNet lexicon (Nimb et al., 2017) assigns at least one frame to 80% of the DDO verbs (and deverbal nouns). The compilation was based on information on thematic and semantic groups in the thesaurus and on valency information from the DDO.

Overall, the strategy for COR-S is:

i) to focus on the core part of the Danish vocabulary in order to guarantee that both the most frequent lemmas and senses but also a large variety of themes and semantic fields are well represented in COR[4] .

ii) to include somewhat simplified versions of the DanNet and FrameNet resources as the semantic part of the lexicon. Simplified in the sense that only a selected number of the most central wordnet and framenet relations and features will be included.

iii) For all the included vocabulary, to reduce the sense inventory adapted from DDO into a core inventory which is more distributionally distinguishable and thus more suitable for automatic processing.

The focus of this paper is on how to semi-automatically accomplish the latter, i.e., to find a principled way to reduce the sense inventory for the entire vocabulary. Our goal is to identify and describe the core word meanings of a lemma and either *delete* the peripheral ones or *merge/cluster* them with their appropriate main sense. We describe in the following the method that we have used to identify such core meanings and subsequently reduce the sense inventory with about 43% percent compared to the DDO.

In Section 2 we describe the linguistic principles of such a reduction and in 3 we present how we develop the gold standard. Section 4 describes our experiments with automatic sense clustering where we experiment with a simple ruled-based model, a word2vec model using cosine similarity to measure the sense proximity, and a deep neural BERT model, which is fine-tuned on the gold standard. Finally, in Section 5 we conclude and outline perspectives for future work.

2. Linguistic principles for a suitable sense granularity level

In 'classical' lexicography and lexical semantics, the discussion of sense granularity has been ongoing for decades, with a typical, slightly simplified, categorisation of lexicographers into being either 'lumpers' or 'splitters' (see among other: Cruse, 1986 and Svensén, 2009). With the rise of NLP, sense inventories have also become an issue in computational lexicography (as discussed in for instance: Fillmore & Atkins, 1992; Kilgarriff, 1997; Agirre & Edmonds, 2007; and Pedersen et al., 2018). Where very rich sense descriptions seem to correspond well to the needs of human users, such very subtle sense descriptions tend to cause notorious problems in automatic processing first of all in automated word sense disambiguation tasks. In fact, this has been the case to an extent where traditional dictionaries have been deemed more or less useless in relation to NLP.

However, several attempts have been made during the years to either manually or semi-automatically adjust sense inventories into coarser-grained ones that are more manageable in NLP. This is shown in for instance Peters et al. (1998), Lapata & Brew (2004), Alvez et al. (2008), Izquierdo et al. (2009), and McCarthy et al. (2016). Most of this work illustrates, however, that such an adjustment is an extremely cumbersome and expensive task and that tools and methods for automating the task are still called for. Most recently, the ELEXIS project (Krek et al., 2018) has developed semi-automated tools for aligning lexicographical datasets by means of semantic text similarity techniques.[5]

[4] We include all DDO lemmas of which at least one sense is linked to one of the PWN core/base concepts via DanNet. Furthermore, we select all DDO lemmas of which at least one sense is a keyword in the thesaurus. The total number of central DDO lemmas identified by this method is approx. 13,000. On top of this, the approx. 60,000 remaining synsets in DanNet will constitute the basis for the selection of supplementary lemmas in the resource.

[5] The NAISC and monolingual dictionary alignment tools, cf. McCrae & Buitelaar 2018, and Martelli et al. 2019 use semantic textual similarity for alignment.

Figure 1: Senses in DDO for the lemma *hær* (army)

Seen from a Danish context, we have previously investigated a number of highly polysemous nouns in the DDO (Pedersen et al., 2018) in order to develop principled ways of reducing the number of senses. In this work, we combine the i) *ontological information* from DanNet with the ii) *main- and sub-sense structure* of the DDO. This combination of features enables us, in a systematic way, to merge/cluster sub-senses of the same ontological type but to maintain i.e. figurative sub-senses of another ontological type.

The existing sense structure in DDO reflects, at least in principle, a close semantic relationship between a main sense and its sub-senses. While sub-senses denote either a broader, a narrower or a figurative nuance of its main sense, main senses are in principle semantically unrelated to each other although etymologically deriving from the same lemma. However, in order to avoid very deep sense structures in the printed dictionary in particular for larger entries (polysemous, typically highly frequent words), senses that in fact could have been classified as sub-senses from the above criteria, are actually sometimes found to be described as main senses. Cf. Figure 1 for the lemma *hær* (army) with two main senses which are actually semantically related to each other (referring to the part of a country's army, which fights on the land, contrary to a country's military power, respectively). This mixed procedure complicates a fully automatic merging of senses. To remedy these inconsistencies and to reduce the number of senses further, we follow the principles of Pedersen et. al. (2018) and *combine* them with a set of additional features. In other words, we supplement the information from DDO on main- and sub-senses with additional lexicographical comments regarding domain-specific use, colloquial use, rare use etc.

In addition, we apply what we have calculated as *sense weight scores* to estimate the 'core-ness' of a particular main or sub-sense. The calculation is based on the amount of lexicographical information attached to the sense in DDO, mainly its number of example sentences, but also the amount of supplementary information, e.g. collocations (the more of these, the more important the sense tends to be). Last but not least, the Danish FrameNet Lexicon is included as a check list for verb and verbal noun senses as well as the status of whether the word is a DT keyword.

With this collection of information types, the following principles have been formulated regarding whether to either delete or merge/cluster senses in COR-S:

Delete a DDO main or sub-sense if it:

- is marked as rare, historic, colloquial, or slang in DDO
- is marked as domain specific in DDO
- generally has a low sense weight score

Merge/cluster a DDO sub-sense with its main sense, unless

- it diverges from the main sense wrt ontological type (from DanNet), i.e. typically concrete ontological types versus abstract types, as is the case of most figurative senses.

All other main senses are maintained – unless they have definitions very close to each other, which is actually found in some cases, Figure 1 being an example of this[6].

A general rule of thumb applied across principles is to pay special attention before merging senses of concrete ontological types like PERSON, ANIMAL or FOOD since we estimate them to be often of particular importance irrespective of their other characteristics.

2.1 Principles for systematic polysemy

Systematic polysemy — i.e. where several lemmas follow the same pattern of polysemy — constitutes its own phenomenon in dictionaries wrt sense structure, and these cases have therefore also evoked specific merge/maintain principles in COR. The way of treating it in DDO proves to rely rather on the frequency of the lemma than on the polysemy type itself[7].

In the COR guidelines, we have collected and classified more than 20 types of systematic polysemy that occur in DDO, encompassing types like PLANT/VEGETABLE, CONTAINER/AMOUNT, BUILDING/INSTITUTION, PROCESS/RESULT etc. For each polysemy type we have decided whether to maintain or merge the involved senses in COR. Frequency can still be included as a decisive factor. However, again the rule of thumb comes into action saying that concrete senses evoke the tendency of wanting to maintain both senses. This means for instance that for

[6] In sum, for *hær* our reduction procedures lead to the following two COR senses: Sense 1: Army/military forces (HUMAN_GROUP) and sense 2: a big quantity of something (ABSTRACT) since the two main senses are semantically close, see also Figure 2.

[7] Frequent words tend to have a more expanded sense structure in DDO than do less frequent lemmas.

process/result where the result is CONCRETE, as in i.e. *bygning* ('building'), we maintain both senses in COR (i.e. both the sense of the act of building as well as the concrete building). In contrast, for process/result where the result is abstract, the principle says to merge, as is the case of *udtalelse* ('saying'/'statement') meaning both the act of saying something as well as the utterance itself.

3. Building the gold standard

The hand-annotated dataset with information on whether to merge/cluster or delete DDO senses when including them in COR, consists of two parts.

The main part contains 3,445 highly polysemous and complex lemmas having at least one sense linked to Princeton WordNet (PWN) via DanNet and corresponding roughly to the so-called core wordnet in PWN (Fellbaum 1998 (ed)). This dataset comprises 2185 nouns (63%), 607 verbs (18%), 472 adjectives (14%), and 119 adverbs (3%). In total the lemmas (3.5% of DDO) have around 15,000 senses (~11% of DDO's senses), even without counting the senses of the many multiword expressions they are part of.

It has been the intention all the way to hand-annotate the most complex part of the vocabulary and not rely on automated methods for this section. 2,148 of the lemmas (62%) are also a DT keyword underlining again their central status.

However, for the training and evaluation of the automatic clustering, there was also the need for a more 'normal' dataset made up of more average vocabulary[8]. Therefore, 2692 polysemous lemmas (1395 nouns, 702 adjectives, and 595 verbs) with between two and five senses and with an average frequency in DDO's corpus, were extracted for manual annotation. Finally, we include DT keywords that are not already included as part of the highly polysemous dataset.

The annotators consisted of a mixed group of trained lexicographers, research assistants and qualified students. The annotators were told to overall follow the above-mentioned principles, but also to *leave room for idiosyncracies*. The idea is to acknowledge the fact experienced by lexicographers that each lemma tends to tell its own semantic story. And further, as mentioned above, that the DDO contains inconsistencies in sense structure here and there which should be adjusted in the process – such as for instance two main senses which are estimated to be so closely related that they should actually be merged. Overall, we achieve a 43% sense reduction, from an average of 4.3 senses in DDO to 2.4 senses in COR.

To check the quality of the principle-driven annotation, a second annotator validated a random 2% subsample of the annotations. Additionally, we validated the most polysemous lemmas that still had seven or more senses after reduction. We use Cohen's *k* to calculate the inter-annotator agreement. The *k* varies between 0.59 and 0.89 with an average agreement of 0.82.

The established sense inventory is further curated in a second round where another annotator also checks the ontological type and hyperonym extracted from DanNet.

4. Experiments with automatic sense clustering

Having hand-coded a subset of the vocabulary – including both complex and average lemmas, the idea is further to automatize the clustering task of the rest of the vocabulary where possible[9]. The overall idea is here to replicate the manual encodings as well as possible by mapping the dictionary information into sense vector representation. Based on the sense representations, we can calculate pairwise sense proximity scores and use the scores as input to our sense clustering algorithm. The task is to partition the set of a lemma's non-deleted senses into *k* subsets that each represents a coarse-grained sense. Here, the gold standard serves for training and evaluation.

As briefly mentioned earlier, monolingual dictionary alignment as performed in the ELEXIS project applies similar techniques (McCrae & Buitelaar, 2018; McCrae et al., 2021). Instead of aligning two monolingual dictionaries as they do, we are essentially aligning a fine-grained dictionary with a coarse-grained version of itself.

However, we face some further challenges when adapting an alignment method for clustering. First and not surprisingly, we have more senses in the source dictionary (i.e., the fine-grained) than in the target dictionary (i.e., the coarse-grained). Therefore, it is necessary to limit the types of links between the source and target, so that multiple source senses can be linked to the same target sense but not vice versa. Secondly, we should expect less overlap in the definitions and quotes than we would when aligning two different dictionaries. When comparing multiple dictionaries, a subset of the sense inventories will overlap and the text in the dictionaries can thereby be similar. Within the same dictionary, the lexicographers work carefully to group and define the various senses and are therefore motivated to write distinct definitions. Thus, the task of reducing through self-alignment increases the challenge of the semantic text similarity calculation (STS). A similar approach to sense clustering is used in the ELEXIS "Clusty"-tool.[10] The key difference is how the senses in the pairs are represented. First of all, the Clusty system presupposes a wordnet sense inventory, while not all our senses have a direct link to a DanNet synset. Secondly, Clusty uses Nasari vectors (Camacho-Collados et al., 2016) which are not currently available for Danish to our knowledge. Instead, we experiment with word2vec and a fine-tuned BERT model for representing senses. Lastly, they use the definitions plus extra information from Wikipedia, where we use all the textual information from the dictionary.

[8] With average we here mean average polysemous.

[9] Note that the task of *deleting* senses is not part of our experiments here since it is a simple task based purely on the unambiguous DDO features.

[10] See https://elex.is/tools-and-services/ and Martelli et al., (2019).

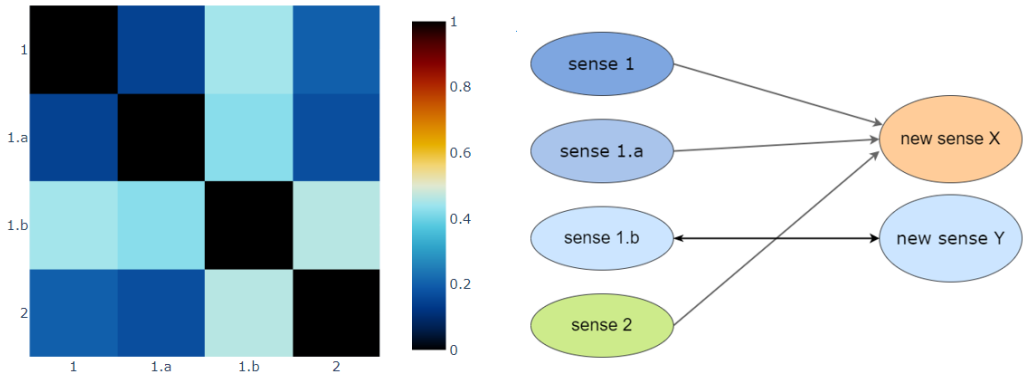


Figure 2: Cosine similarity of the centroid word2vec embedding of the definitions compared to each other for the Danish word *hær* 'army'. *To the right:* the allocation of the DDO senses into broader clusters in COR.

4.1 Dictionary information for sense representations

To represent the senses, we experiment with both a feature-based and a text-based approach. The feature-based approach uses the non-textual information in the dictionary, which includes the sense hierarchy, DanNet ontology, and figurative markings. The features constitute be the input for a simple rule-based model, which follows the principles listed in Section 2 and the experiment in Pedersen et al. (2018). However, this approach may not capture the cases where the annotators deviate from the principles and thus, the clusters may not correspond to the core of Danish senses as a human would perceive them.

In early experiments, we found that the information in the feature representations led to a bias towards never merging two main senses, while the human annotators do merge two main senses if their definitions and quotes are similar. In the cases of figurative senses, we expect the definitions and quotes to be distinct enough from the literal senses for a text-based system to differentiate them. Therefore, we also experiment with a text-based approach in which we apply the textual input directly available from DDO and thus exclude the additional features. The textual information can be vectorized straight-forwardly, and the definitions and quotes are available for almost all senses[11] (whereas DanNet features are only available for a part of the vocabulary since not all DDO senses are in DanNet).

We estimate the ability to merge two senses by calculating a sense proximity score. Since we are primarily working with dictionary text in our sense representations, we can also view the estimation as an STS problem (Agirre et al., 2016). However, where the typical STS task returns a similarity score between 0-5 (unrelated to complete equivalence), we only calculate a single similarity score between 0-1 that estimates how much meaning the two senses share. We can view the sense representations as creating a semantic vector space where each dimension corresponds to a piece of information. If two senses are closely located in the semantic space, then we can assume they share some meaning. Therefore, we can use distance measures like the cosine similarity measure to estimate the sense proximity score. This approach is used in combination with a Danish word2vec model (Sørensen & Nimb, 2018). Here, we will embed every content word in the textual information of a sense and then calculate the centroid embedding. All the sense embeddings for a lemma are then compared.

An example of this is seen in Figure 2, where we have plotted a cosine similarity matrix of the senses of the aforementioned Danish noun *hær* ('army'). All the senses except sense 1b mean something related to the military forces, while 1b is a figurative sense meaning 'a large number' (for the full set of sense definitions, see Figure 1). According to the calculated proximity score, the two main senses 1 and 2 can be merged and the figurative sense should be maintained. This is in accordance with the human annotations, even if the merging of main senses would cause problems for a feature-based approach.

Another method for estimating the sense proximity is to use the sense representations as input to a classification model. We train the model using our annotated dataset by inferring the labels from the manual clusters. The advantage of this approach is that we can fit the model to the sense granularity of our choice – in our case the hand-coded examples. Instead of outputting the actual label, we use the probability output as a proximity score. In this way, we can take the uncertainty of the model into account in the allocation algorithm. We experiment with this second text-based method using a Danish pretrained BERT, which can be fine-tuned to perform a sentence classification task for the STS purpose as described in McCrae et al (2021).

The clustering algorithm allocates each sense to a cluster based on the previously calculated sense proximity score. The task is fundamentally the same as aligning two sense inventories based on pairwise similarity scores. However, we put further constraints on the alignment since we specifically aim to produce a coarser grained version of the same dictionary. Therefore, we only allow a many-to-one (alignment to broader sense) and one-to-one (exact sense alignment) mapping.

Following the human annotations, in some cases, more than two senses are merged. To allocate senses to a broader

[11] Out of all the 15 000 senses, 300 were without a quote. In these cases, we tried to assign a suitable quote.

COR sense, multiple pair scores must be considered in the decision. For instance, if a lemma has a set of senses: *A = {a, b, c, d}*, which is further partitioned into subsets by the human annotators: *X = {a, b, d}, Y = {c}*. To recreate the merged sense *X*, the proximity scores between all the members of *X* must be high. Additionally, all the proximity scores between a member of X and a member Y must be low; see Figure 2. The goal of the clustering is therefore to correctly assign the senses to a cluster based on all the proximity scores for a lemma. Thus, we evaluate the clustering by comparing the partitions in the gold standard with the partitions of the automatic methods using the Rand index (RI) (Rand, 1971).

The actual alignment is performed by a simple algorithm, which either merges or splits a sense pair based on threshold values. The thresholds are found from analysing the distribution of mergeability scores across the two labels in a development set. We repeat the tuning of the algorithm for every model. The senses of each lemma are processed as a group with a stack of pairs to be processed. The stack of pairs is sorted by how certain the merging classifier is. If a pair has received a very high or a very low score, then it is assumed these pairs are correctly classified in the previous step. If a pair falls into an area of uncertainty, then the alignment is based partially on the current alignments in the group and partially on whether the score falls closest to the group's maximum or minimum. The last criterion is driven by a pattern found in the mergeability scores.

4.2 The models

We experiment with three different models: i) a simple ruled-based model that follows our principles from Section 2, ii) a word2vec model using cosine similarity to measure the sense proximity, and iii) a deep neural model in form of a BERT model that is fine-tuned on the annotations.

The *rule-based model* classifies a sense pair on basis of three criteria: i) Whether it is part of the same main sense, ii) whether one is figurative, and iii) whether they are assigned the same ontological type (if available). If the ontological type is not available for both senses, then it is assumed that the senses have the same ontological type.[12]

The word2vec model (Mikolov et al., 2013) is a 500-dimension skipgram model trained by DSL (Sørensen & Nimb, 2018). The motivation for using a word2vec is the fact that the model we are using is trained on the DSL text corpus - the same corpus used for creating and updating the fine-grained dictionary. The word2vec model is also used in a tool for finding related words to a sense. Therefore, we imagine the model could be useful for this context as well. However, it still has the problem of conflated senses (Camacho-Collados & Pilehvar, 2018), that is, each representation is based on a lemma form and not a sense. In order to adapt the word2vec model to represent senses, we follow the method of Olsen et al. (2020) and create a combined word2vec from all the content words in the textual data for a specific sense. That is, we calculate the centroid embedding of the set of words in the textual data after stop words and special characters are removed. However, we do not use the DanNet relations as not all senses had a corresponding DanNet synset. The resulting sense embeddings create a semantic space in which each sense is a vector. The semantic proximity can then be measured by the cosine distance measure.

To experiment with *contextualized embeddings*, we apply a BERT model (Devlin et al., 2019) which has the advantage of allowing us to fine-tune the model to a sentence classification task. Since a contextualised embedding represents a given lemma in a specific context, we can assume that the embeddings to some extent resembles the senses. We map the BERT representations onto our sense inventory by fine-tuning the BERT using a dataset automatically compiled from the annotations. The fine-tuning simply adds a classification layer on top of the model that can be trained on a specific task – in our case, a sentence classification task. An advantage of the fine-tuning is that we can use the output of the fine-tuning layer as a semantic proximity score, thereby surpassing the need for a metric like the cosine.

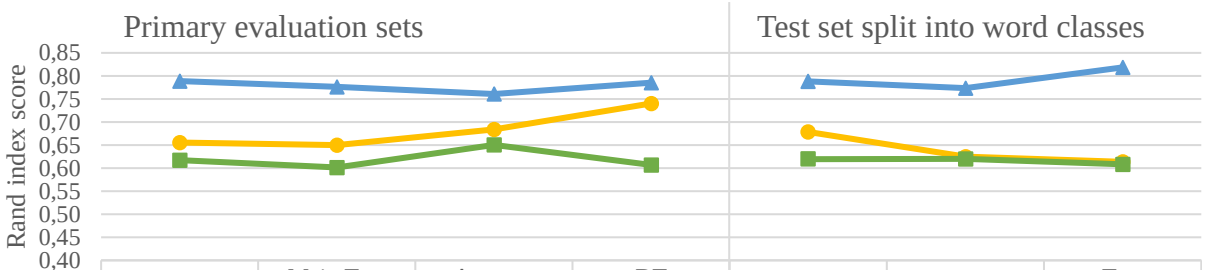
The model we use is the publicly available pretrained BERT from Certainly[13]. This model is trained on 1.6 billion Danish lowercased words from a corpus compiled from Common Crawl, Danish OpenSubtitles (Lison & Tiedemann, 2016), Danish Wikipedia, and web scraped forums.

Our method for fine-tuning the BERT combines the Gloss Selection Objective (GSO) used for Word Sense Disambiguation with a BERT (Huang et al, 2019; Yap et al., 2020) and the Word-in-Context dataset (WiC) (Pilehvar & Camacho-Collados, 2018). The GSO utilises the inherent sequence pair architecture of BERT to compare senses as seen in specific contexts (e.g., the context vs. gloss). The purpose of this task is to select the most relevant context-gloss pair from a list of related pairs. The context is a sentence containing a target lemma, while the gloss is a sentence from a lexical resource (e.g., a wordnet) representing a specific sense of the target lemma. The relevant-pairs list is constructed by pairing the same context sentence with a number of possible glosses (i.e., senses). The model compares all the context-gloss pairs, and the highest scoring pair is chosen as the best match. The sense that is represented as the gloss in the best match is therefore chosen as the sense of the target lemma in the context sentence.

We use a different method for retrieving context sentences in our dataset compared to the original GSO. In Yap et al. (2020), they use the sense-tagged dataset semcor (Miller et al., 1994) and the English wordnet (Yap et al., 2020) to retrieve the contexts and glosses, respectively. For Danish, there are only few and small sense-tagged datasets. For instance, the SemDaX lexical sample dataset (Pedersen et al., 2016) only tags the 20 most polysemous nouns. Instead, we use the compiling method of WiC which is constructed by pairing the quotes from a wordnet that shares a target lemma. The sense of the target is inferred from the wordnet

[12] DDO contains more senses than DanNet and therefore not every DDO sense has an ontological type. We assume that non-figurate senses have the same ontological type within a main sense unless other is otherwise specified.

[13] https://github.com/certainlyio/nordic_bert



	Main Test	Main Test Reduced	Average Vocab	DT keywords	Test Nouns	Test Verbs	Test Adjectives
rule-based	0,79	0,78	0,76	0,79	0,79	0,77	0,82
BERT	0,66	0,65	0,68	0,74	0,68	0,62	0,61
word2vec	0,62	0,60	0,65	0,61	0,62	0,62	0,61

Figure 3: The RI clustering scores across the four primary evaluation sets (left) and across the test set split after word classes (right)

synset that the quote originated from. The WiC task is then for a model to decide whether two quotes with the same target contains the same sense or two different senses. This task is essentially the same as an isolated comparison of context-gloss in the GSO, but where the context is also retrieved from a lexical resource.

We compiled a fine-tuning dataset from the gold standard and additional quotes from the Danish dictionary. This dataset is then used to improve the BERT model's pairwise semantic comparison abilities. Each pair (context-gloss) in the relevant-pairs list represents a COR-S sense. The task is to find the pair where both the context and gloss comes from the same COR-S sense. We also add an extra binary task of classifying whether a sentence pair belongs to the same sense.

The context-gloss pair is either a quote-definition or definition-definition pair. The definition-definition pairs are only possible for those senses that are the result of merging multiple fine-grained senses in the annotation. To ensure the model is aware of the position of the target lemma, we inserted a [TGT]-marker on both sides of the target. Since the definitions do not include the target, we prepend the target along with the markers. The sentences are also further preprocessed by lowercasing and removing special characters.

In total from 2275 polysemous lemmas with a total of 7026 senses, we achieved 16 230 training instances. The BERT model is fine-tuned on 80% of this compiled fine-tuning dataset with a learning rate of 10^{-5} over 2 epochs (20 204 steps). The batch size varies depending on the number of senses for each lemma.

4. 3 The datasets

The annotation is split into several subsets for training, development, and testing. These subsets are divided based on number of lemmas, since each lemma should be processed as a group of senses. The training set is composed of 80% of the main, highly polysemous dataset. This set is used for the fine-tuning of BERT. The rest of the main dataset is split into two sets of 10% for development and testing, respectively. The development set is used for setting the hyperparameters of the BERT fine-tuning and the thresholds for the clustering algorithms. Additionally, we split the DT keywords and average vocabulary datasets into a training and test set. In the experiments, we only use the test sets, and we save the training sets for ongoing work. We decided to keep a large proportion of the data as a test set (50%) to give a better estimate of the possibility to expand the clustering to the rest of the vocabulary. Likewise, we create a light version of the main test set by removing any lemma with more than 5 senses to better understand the possible negative influence of the difficult, highly polysemous lemmas. Furthermore, we are interested in the performance on the different word classes, and we split the main test set according to word classes[14]. In total, we end up with four primary evaluation subsets (test, average vocabulary, DT keywords, test reduced), and three word-class subsets (nouns, verbs, adjectives).

4.4 Results

We report the clustering results in Figure 3. As can be seen, the rule-based model outperforms the text-based models on every evaluation subset. It does particularly well on the adjectives, achieving a high membership score of 0.82. The model achieves a similar RI score on the other subsets, though slightly worse on the average vocabulary and verbs. The RI score of 0.79 on the test set is in particular impressive considering that subset contains the most polysemous lemmas and is assumed to be the most difficult set.

The BERT model performs the best out of the text-based models. Though it never manages to outperform the rule-based model, the gap is closer on the average vocabulary and DT keywords sets. This confirms our decision of hand-coding the most challenging lemmas of the vocabulary. The difference in performance of the rule-based model and BERT on the main and average vocabulary can be explained by the sense hierarchy of the two sets. Since the average vocabulary has a lower number of average senses, the set contains less subsenses. Unlike the rule-based model, BERT does not rely on the sense hierarchies and is therefore better at estimating when to cluster pairs in the less polysemous subset. Therefore, it might be possible to improve our method by combining the rule-based approach and the BERT by adding the BERT score to the feature input vector used for the rule-based model. A classifier

[14] Note that adverbs are not included.

could then be trained on these feature vectors, which in turn could output more robust sense proximity scores that allows for main sense clustering.

The word2vec model performs the overall worst out of the three models. When we inspect the distribution of the sense proximity scores over the clustering labels (e.g., whether a sense pair is merged in the annotation), we do see a difference between the labels, though there is a considerable area of overlap in the proximity scores. The model can capture and compare some of the fine-grained dictionary senses, though not consistently enough to make it useful for our purpose. Surprisingly, the performance is not aided by removing the most polysemous lemmas in the reduced test set. In contrary, the model performs the best on the average vocab set that also only contains lemmas with fewer than five senses. It appears that the information in the definitions and quotes are not enough to make distinguishable sense representation using a word2vec model. Informing the model with additional information may be a good way to proceed.

The results on word classes show that the verbs are more difficult for the rule-based and BERT model. We explain the difference with the complex sense structure some verbs have. For instance, the verb *stå* 'stand' has over 30 senses distributed over 11 sub-senses in DDO. Adjectives also introduce a higher challenge for BERT, but the rule-based model thrives in this condition. One can explain this difference by the fact that adjectives typically have many sub-senses and similar ontological types in DanNet, which favours the rule-based model. With the differing performance on word classes, it should investigated whether to develop separate approaches for each word class.

It can be questioned whether the fine-tuning may have a negative influence on the BERT model's ability to estimate sense proximity. Since the model is presented with one positive and n-senses negative examples for each batch, it may be somewhat biased from the higher number of negative cases. In an attempt to remedy this, we re-tuned the BERT with a reduced number of negative senses and without the GSO, however without improving the clustering.

Another problem may arise from the clustering BERT setup. Since more than one sentence pair can be found for a sense pair (e.g., definition-quote, definition-definition), the algorithm has multiple proximity scores for the same sense pair. The algorithm trusts the most probable scores first (very high and very low scores), however if both scores are probable and contradicting, the algorithm may choose to incorrectly split or cluster a sense pair. It should be considered in future work if there are better solutions for the clustering algorithm to make it more robust to disagreements between sense pairs.

5. Conclusions and future work

In this paper we have examined the notion of sense granularity with the purpose of developing a new lexicographical resource for AI purposes. The overall idea is to take advantage of the very rich and socially contextualised information on word meaning already described in traditional lexica, like the DDO. We introduce a concept of 'core-ness' and outline a number of principles of how to achieve such a core sense inventory. The aim is to compile a sense inventory that is more practically useful for contemporary text material and more directly suitable for AI, i.e., omitting outdated language and slang, merging subtle and rare sub-senses with their main sense, disregarding sub-senses that belong to very specific domains, etc.

As we have shown, a substantial part of the vocabulary has undergone manual simplification according to these lexicographical principles resulting in an extensive gold standard of more than 6,000 lemmas where senses are reduced for COR with approx. 43%. These comprise both very polysemous lemmas and more average words. The intercoder agreement for this work is relatively high with an average agreement of 0.82 using Cohen's *k*.

Our results on automatic sense clustering are promising at least for a subset of the vocabulary. Our experiments indicate that the rule-based model currently provides the best results per se, in particular on adjectives, even if also the BERT model looks promising. Generally, we note that the sense clustering task is overall quite challenging, especially for a lower-resourced language like Danish where data is still somewhat sparse (wrt. i.e. sense annotated text resources) and maybe to some extent even biased towards the more complex part of the vocabulary. In our case, as we have seen, the main training and test data is comprised by a high number of very polysemous lemmas, and for some of this data, even the human annotators have difficulties in agreeing on how to merge senses. This goes in particular for the verbs.

Overall, we have shown that removing the most polysemous lemmas from the dataset increases the robustness of the automatic text-based methods. This speaks in favour of continuing to hand-code the most complex part of the vocabulary leaving only the more average lemmas for automatic reduction.

Further, we should consider in the future whether to work more individually with the three word classes when training our models, since they tend to perform quite differently wrt. sense structure.

Acknowledgements

The COR development project is funded by the Danish Agency for Digitisation as part of an AI initiative embarked by the Danish Government in 2020. The research behind the COR project, however, relies substantially on the European Lexicographic Infrastructure (ELEXIS) project under the European Union's Horizon 2020 research and innovation programme (grant agreement No 731015).

Bibliographical references

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez Agirre, A., Mihalcea, R., & Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511*. ACL (Association for Computational Linguistics).

Agirre, E., & Edmonds, P. (Eds.). (2007). Word sense disambiguation: Algorithms and applications (Vol. 33). Springer Science & Business Media.

Alvez, J., Atserias, J., . Carrera J., Climent, S., Laparra, E., Oliver, A.& Rigau, G. (2008). Complete and consistent annotation of wordnet using the top concept ontology. *LREC Proceedings 2008*.

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, *240*, 36-64.

Camacho-Collados, José, & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, *63*, 743-788.

Cruse, D. A. (1986). Lexical semantics. Cambridge university press.

Devlin, J., Chang, M-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics

Fellbaum, C. (ed.). (1998). WordNet: An Electronic Lexical Database. MIT press.

Fillmore, C. J., & Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. Frames, fields and contrasts: New essays in semantic and lexical organization, 75, 102.

Huang, L., Sun, C., Qiu, X., & Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Izquirdo, R., Suárez, A. & Rigau, G. (2009). An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 389-397. The Association for Computational Linguistics.

Kilgarriff, A. (1997). I don't believe in word senses. Computers and the Humanities, 31(2), 91-113.

Kosem, I., Nimb, S., Tiberius, C., Boelhouwer, B.& Krek, S. (2021). License to use: ELEXIS survey on licensing lexicographic data and software. In Mitits, Lydia and Kiosses, Sypros (eds.) *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7-9 September 2021, Alexandroupolis, Vol. 2*. Democritus University of Thrace, pp. 705--712

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen B. S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana, Slovenia.

Kirchmeier, S., Henrichsen P. J., Diderichsen, P. & Hansen, N. B. (2019). *Dansk Sprogteknologi i Verdensklasse*. Dansk Sprognævn, https://dsn.dk/wp-content/uploads/2021/01/sprogteknologi-i-verdensklasse.pdf.

Lapata, M. & Brew, C. (2004). Verb Class Disamsubstantially biguation Using Informative Priors. *Computational Linguistics*, 30(1): 45-73.

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Martelli, F., Navigli, R., Spadoni, P., Stilo, G. & Velardi, P. (2019). Lexical-Semantic Analytics for NLP: Sense Clustering, Project Report, D3.1, ELEXIS. https://elex.is/wp-content/uploads/2019/08/ELEXIS_D3_1_Lexical_semantic_analytics_for_NLP_sense_clustering_Final.pdf

McCrae, J. P., Ahmadi, S., Yim, S. B., & Bajčetić, L. (2021). The ELEXIS System for Monolingual Sense Linking in Dictionaries. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, 68.

McCrae, J. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. Cybernetics and Information Technologies, 18(1):109–123.

Mikolov, T., Yih, W. & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746-751).

Miller, G., Chodorow, M., Landes, S., Leacock, C. & Thomas, R. (1994). Using a Semantic Concordance for Sense Identification. In *Proceedings of the ARPA Human Language Technology Workshop*. 10.3115/1075812.1075866.

Nimb, S. (2018). The Danish FrameNet Lexicon: Method and lexical coverage. In: *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructions*, 48–52.

Nimb, S., Braasch, A., Olsen, S., Pedersen, B. S. & Søgaard, A. (2017). From Thesaurus to FrameNet. I: Proceedings of eLex 2017. Electronic Lexicography in the 21st century - Proceedings of eLex 2017 conference. s. 1-22. Leiden, The Netherlands.

Nimb, S., Lorentzen, H., Theilgaard, L. & Troelsgård, T. (2014 a). Den Danske Begrebsordbog, Det Danske Sprog- og Litteraturselskab, Copenhagen, Denmark.

Nimb, S., Lorentzen H. & Trap-Jensen, L (2014 b). The Danish Thesaurus: Problems and Perspectives In: Abel A., Vettori C. & Ralli N. (eds.). *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen 2014: EURAC Research, pp. 191-199.

Olsen, IR, Sayeed, A & Pedersen, B. S. (2020). Building Sense Representations in Danish by Combining Word Embeddings with Lexical Resources. In *Globalex Workshop on Linked Lexicography: LREC 2020 Workshop Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, s. 45-52. <https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/GLOBALEX2020book.pdf>

Pedersen B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L. & Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing

a monolingual dictionary. In Language Resources and Evaluation, 43, (pp. 269-299).

Pedersen, B. S., Braasch, A., Johannsen, A. T., Martinez Alonso, H., Nimb, S., Olsen, S., Søgaard, A., & Sørensen, N. H. (2016). The SemDaX Corpus - sense annotations with scalable sense inventories. In *Proceedings of the 10th conference of the Language Resources and Evaluation Conference* (pp. 842-847). European Language Resources Association.

Pedersen, B. S., Aguirrezabal Zabaleta, M., Nimb, S., Olsen, S., & Rørmann, I. (2018). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In Proceedings of Global WordNet Conference 2018 Global WordNet Association. http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/gwc-2018-proceedings.pdf

Pedersen, B., Nimb, S., Olsen, I. R. & Olsen, S. (2019). Linking DanNet with Princeton WordNet. In *Global WordNet 2019 Proceedings*, Wroclaw, Poland Oficyna Wydawnicza Politechniki Wroclawskiej.

Peters, W., Peters, Y. & Vossen, P. (1998.). Automatic sense clustering in EuroWordNet. In: First International Conference on Language Resources & Evaluation 1998, Granada, Spain.

Pilehvar, M. T., & Camacho-Collados, J. (2018). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXivpreprint arXiv:1808.09121*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66* (336), 846-850.

Svensén, B. (2009) A Handbook of Lexicography. Cambridge University Press.

Sørensen, N. H., & Nimb, S. (2018). Word2Dict-Lemma Selection and Dictionary Editing Assisted by Word Embeddings. In *Proceedings of the 18th EURALEX International Congres: Lexocography in Global Contexts* (pp. 819-827).

Yap, B. P., Koh, A., & Chng, E. S. (2020). Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences. *arXiv preprint arXiv:2009.11795*.

Language Resource References

DanNet : https://cst.ku.dk/english/projects/dannet/

Den Danske Ordbog (DDO): Hjorth, E. & K. Kristensen red. (2003-2005). *Den Danske Ordbog*, volume 1-6, Det danske Sprog- og Litteraturselskab/Gyldendal, Copenhagen. Online: ordnet.dk/ddo

Den Danske Begrebsordbog (DDB): Nimb, Sanni, Henrik Lorentzen, Thomas Troelsgård, Liisa Theilgaard, Lars Trap-Jensen (2014). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab, København, Danmark.

The Danish FrameNet Lexicon: https://korpus.dsl.dk/resources/details/framenet.html#english.