Proceedings of the Globalex Workshop on Linked Lexicography, pages 45–52
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Building Sense Representations in Danish by Combining Word Embeddings with Lexical Resources

Ida Rørmann Olsen[1], Bolette S. Pedersen[2], Asad Sayeed[3]
Centre for Language Technology, University of Copenhagen[1,2],
Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg[3]
Emil Holms Kanal 2, 2300 Kbh S[1,2], Renströmsgatan 6, 412 55 Gothenburg[3]
idaroermannolsen@gmail.com, bspedersen@hum.ku.dk, asad.sayeed@gu.se

Abstract

Our aim is to identify suitable sense representations for NLP in Danish. We investigate sense inventories that correlate with human interpretations of word meaning and ambiguity as typically described in dictionaries and wordnets *and* that are well reflected distributionally as expressed in word embeddings. To this end, we study a number of highly ambiguous Danish nouns and examine the effectiveness of sense representations constructed by combining vectors from a distributional model with the information from a wordnet. We establish representations based on centroids obtained from wordnet synsets and example sentences as well as representations established via a clustering approach; these representations are tested in a word sense disambiguation task. We conclude that the more information extracted from the wordnet entries (example sentence, definition, semantic relations) the more successful the sense representation vector.

Keywords: Danish, wordnet embeddings, word sense disambiguation

1. Introduction

The effective handling of sense ambiguity in Natural Language Processing (NLP) is an extremely challenging task, as is well described in the literature (Kilgarriff, 1997; Agirre and Edmonds, 2006; Palmer et al., 2004; Navigli and Di Marco, 2013; Edmonds and Kilgarriff, 2002; Mihalcea et al., 2004; Pradhan et al., 2007).

In this paper, we focus on a lower-resourced language, Danish, with the hypothesis that if we can compile sense inventories that *both* correlate well with human interpretations of word meaning *and* are well-reflected statistically in large corpora, we would have made a first and important step towards an improved and useful sense inventory: not too fine-grained, but still capturing the essential meaning differences that are relevant in language processing. We investigate this hypothesis by building sense representations from word embeddings using wordnet-associated data.

In order to assess the performance of the proposed model, we study a number of Danish nouns with very high meaning complexity, i.e., nouns that are described in lexica as being *extremely* polysemous. We apply a central semantic NLP task as our test scenario, namely that of *word sense disambiguation* (WSD). For lower-resourced languages, obtaining performance better than a majority-class baseline in WSD tasks is very difficult due to the extremely unbalanced distribution of senses in small corpora. However, the task is an ideal platform for achieving our goal of examining different approaches to sense representation. Our aim is both to support a data-driven basis for distinguishing between senses when compiling new lexical resources and also to enrich and supplement our lexical resource with distributional information from the word embedding model.

In the following, we carry out a series of experiments and evaluate the sense representations in a WSD lexical sample task. For the experiments, we represent wordnet synset information from the Danish wordnet, DanNet (Pedersen et al., 2009), in a word embedding model. We test five different Bag-Of-Words (BOWs) combinations—defined as 'sense-bags'—that we derive from the synsets, including information such as example sentence, definition, and semantic relations. Generally speaking, the synsets incorporate associated concepts via semantic relations which lexicographers have chosen as being the defining relation for each particular concept. This approach sheds light on the extent to which the hand-picked words in the synsets are actually representative of the processed corpus data.

It is not possible at this stage to evaluate an unsupervised word sense induction (WSI) system for Danish with curated open-source data. However, with a knowledge-based system, where the sense representations are linked to lexical entries, it is possible to evaluate with the semantically annotated data available for Danish, the SemDaX Corpus (Pedersen et al., 2016). This corpus is annotated with dictionary senses.

The paper is structured as follows: Section 2 describes Danish as a lower-resourced language and presents existing semantic resources that are available for our task. In Section 3, we present related work, and in Section 4 we describe our five experiments in detail. Section 5 and 6 describe and discuss our results, and in Section 7 we conclude and outline plans for future work.

2. Danish as a lower-resourced language

Semantic processing of lower-resourced languages is a challenging enterprise typically calling for combined methods of applying both supervised and unsupervised methods in combination with language transfer from richer-resourced languages. For Danish we have now a number of standard semantic resources and tools such as a wordnet and SemDaX corpus, a framenet lexicon (Pedersen et al., 2018b), several word embedding models (Sørensen and Nimb, 2018), and a preliminary sense tagger (Martinez Alonso et al., 2015). However, the size and accessibility of the resources as well as the evaluation datasets accompanying them typically constitute a bottleneck.

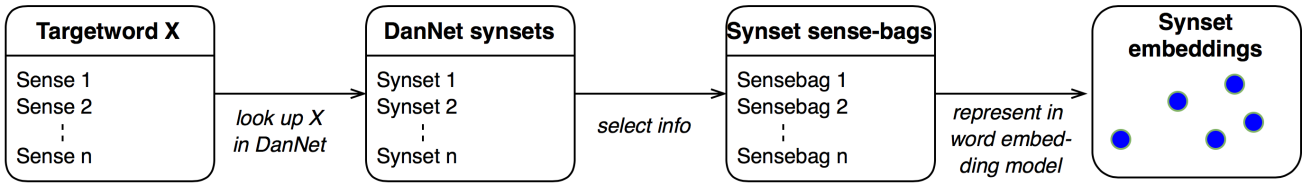

Figure 1: The method used to build the synset embeddings.

For instance, the wordnet, DanNet, which contains 65,000 synsets, is open-source, but the links from DanNet to the complete sense inventory of The Danish Dictionary is not. Our work requires this key, which necessitated connecting the dictionary labels to DanNet synsets through cumbersome manual compilation.[1]

3. Related Work

Both supervised and unsupervised methods to represent words and word senses have been widely explored in NLP, especially given the popularity of word embeddings. Unsupervised approaches to obtain not only word embeddings, but also sense embeddings (such as SenseGram (Pelevina et al., 2017), Adagram (Bartunov et al., 2016), and Neelakantan et al. (2014)) do not rely on existing large datasets; they are are thus suitable for lower-resourced languages. A downside is that the induced senses are not humanly readable or easy to link to lexical resources; this limits their applicability.

An incorporation of valuable high-quality resources, e.g., wordnets, in unsupervised methods can augment the sense representations with additional lexical information, especially for non-frequent word senses. The combination of contextual and knowledge-based information can be established by joint training (Faralli et al., 2016; Johansson and Nieto-Piña, 2015; Mancini et al., 2017), or by post-processing normal word embeddings (Rothe and Schütze, 2017; Bhingardive et al., 2015; Chen et al., 2014; Pilehvar and Collier, 2016; Camacho-Collados et al., 2016). Alternatively, Saedi et al. (2018) successfully converted a semantic network (WordNet) into a semantic space, where the semantic affinity of two words is stronger when they are closer in the semantic network (in terms of paths). They tested the resulting representations in a semantic similarity task and found a significant improvement compared to a regular word2vec space. The study also indicated that the more semantic relations included from the semantic network, the better the result.

Bhingardive et al. (2015) detected the most frequent senses by comparing the target word embedding in a word embedding model with constructed sense representations based on synset information represented in a word embedding model. Our work is also related to Ustalov et al. (2018) who proposed a synset-averaged sense-embedding approach to WSD for an under-resourced language (Russian). They evaluate the system's clustering on a gold-standard with an average number of word senses of 3.2 (Panchenko et al., 2018). Their results show that the task of building unsupervised sense embeddings this way is remarkably difficult.

We estimate the quality of the sense representations in a lexical sample WSD task. The contribution of this paper is therefore a study on these methods for Danish data evaluated on a WSD task and not for most frequent sense detection or on a gold standard. The work provides a detailed investigation of which information types from DanNet improve our WSD results, and with more focus on the role of example sentences than seen in related work.

4. Five word embedding experiments

For a number of years up to now, embeddings have been ubiquitous in computational approaches to numerous NLP tasks. While word embeddings, such as word2vec (Mikolov et al., 2013), have been central in NLP research touching on lexical semantics, other forms of embeddings, from character to paragraph to multimodal, have proven to be flexible, often multi-purpose forms of linguistic representation. Our overall idea is to build sense representations in vector spaces with information of associated words extracted from a lexical resource, namely wordnet. We make use of word embeddings to construct a sense representation, a *synset embedding*. The wordnet synset information (i.e., words) associated to a given sense of a word is collected in a synset "sense-bag". The synset sense-bag is used to construct a unified sense representation, the *synset embedding*, inside a word embedding model. See Figure 1.

Note that for each synset, DanNet provides both the handpicked related concepts (as illustrated in Figure 2), one handpicked example sentence where the sense is used in context, and (part of) the sense definition from The Danish Dictionary.

For example, a particular synset sense-bag of the polysemous Danish targetword *model* (approximately the same concept as in English)—in the sense of a representation of something (sometimes on a smaller scale) consists of the example sentence: *"Færgen er en model i 1:4"* and the synset members *Effekt, videnskab, fremstille, figur, afprøve, gengive, pynte, arbejdsmodel, gine, globus, globus, mockup, modelbygning, modelfly, skalamodel, skibsmodel, modeljernbane, modelbil, modelskib, modeltog, kirkeskib* [2].

[1] We build the sense representations with DanNet, but our evaluation data, SemDaX, is annotated with dictionary labels. The Danish Dictionary is not fully available for research.

[2] "The ferry is a model 1:4", Effect, science, produce, figure, test, represent, decorate, working model, gine, globus, mock-up, model building, airplane model, scale model, ship model, train-track model, car model, ship model, train model, church ship

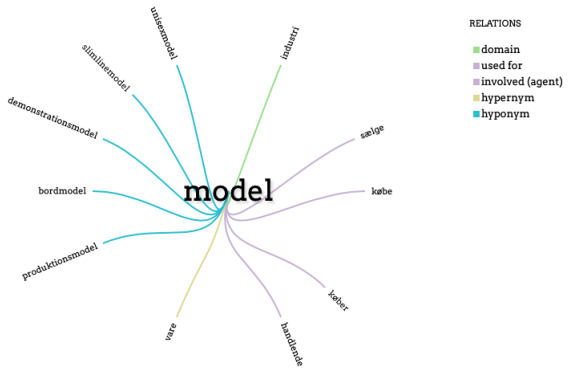

Figure 2: A synset of the targetword *model* (as in a model in industrial production). Semantic relations on the right.

In addition, the synset sense-bag of *model* in the sense of a schematic description or illustration of an abstract, complicated thing or relation, has the example sentence *"Watson og Crick fremsatte deres model af DNA-molekylet som en dobbeltspiral, der kan visualiseres som en vredet stige"* and the synset members *Anskueliggørelse, videnskab, atommodel, forklaringsmodel*[3].

First, we construct synset embeddings represented in a word embedding model by unifying information extracted from DanNet for each sense of the target nouns. These synset embeddings are tested in a WSD task using cosine similarity. Second, we apply the synset embeddings to sense-tag new unannotated data via a clustering approach. By doing this, we build more corpus-influenced synset embeddings (i.e. synset embeddings not exclusively built from wordnet information) and, at the same time, also obtain training data of a proper size to benefit of the advantages of machine learning models for future WSD experiments. See details of the method in section 4.1.

The method will work when there is a correspondence between how words in the knowledge-base for the given lexical resource (DanNet) are distributed across senses and what the distributional information of the words looks like in the word embedding model. If the words associated for each sense in DanNet are important for the concept's use in language, then the collection of those words in the word embedding model is reasonable, since such a model represents word similarity based on the distribution of words used in data.

The approach can be seen as highly scalable since the sense representations can be obtained without full annotation of a training corpus and is applicable for all word entries included in the input resource. The method would therefore be applicable also to other lower-resourced languages.

It should be emphasized that we test our approach *both* on a set of some of the most polysemous words found in Danish *and* operate on the most fine-grained version of the applied evaluation data (the SemDaX Corpus). Working with this corpus, Pedersen et al. (2018a) suggested a principled approach to sense clustering. In that work, the coarsest sense granularity level proved to be most operational (in a WSD task), obtaining the highest inter-annotator agreement score. In our work, however, we choose the finest level of granularity to access the potential of the method when tested on a really hard task.

[3] *"Watson and Crick presented their model of a DNA molecule like a double-spiral, that can be visualized like a twisted latter", visualization, science, atom model, explanation model.*

4.1. Experiment Details

We collect various synset information in synset sense-bags, and each word sense representation (synset embedding) is the centroid of the word embeddings from the corresponding sense-bag. The word embeddings originate from the word2vec word embedding model described in section 4.3., and the constructed synset embeddings live within that same vector space. The synset information varies for each experiment.

More precisely, a synset sense-bag is a set, $B = \{w_1, \ldots w_n\}$, where n is the number of words in B and the w's are the words selected[4] from the synset information. Each word, w_i in B, can be represented by a word vector $\overrightarrow{W_i}$ in the word embedding model. These word vectors in B are averaged into a mean vector, $\overrightarrow{M}$, where $\overrightarrow{M} = \frac{\sum \overrightarrow{W_i}}{n}$. $\overrightarrow{M}$ is the resulting synset embedding of the given synset sense-bag, B. Therefore, for each sense of each targetword we can collect a synset sense-bag, B, from DanNet and construct a synset embedding, $\overrightarrow{M}$, with the word embedding model. The extracted information from DanNet contain only words (not numbers). The words are not weighted when constructing the synset embedding with their word embeddings. Multi-word terms are treated as multiple words under word tokenization (these instances are rarer in Danish, than in English). In doing this, we examine whether the selected knowledge-based information from DanNet in combination with the distributional representation of the words in the synset sense-bags can construct appropriate sense representations.

Four types of synset sense-bags for building synset embeddings are tested:

1. Local synset members: Collection of hypernyms, hyponyms, synonyms, near-synonyms, used-for and made-by semantic relations, together with the bag-of-words (BOW) of the word sense definition.
2. Example sentence: BOW from the example sentence using the sense in context.
3. Example sentence+: BOW collection of local raw example sentence *and* raw example sentences from the hypo- and hypernym synsets.
4. Combination: All collections from exp. 1-3 put together *and* the BOW of definitions of hypo- and hypernyms.

A fifth and final synset embedding is tested, in which the best performing synset embedding above is used as a seed in the k-means algorithm (Lloyd, 1982) to auto-tag unannotated example context sentences by a clustering approach:

[4] Selected according to the given experiment.

5. Cluster centroid: Centroid of clustered context vectors

The idea is to tune the synset embeddings by adding more data than merely information from DanNet. The seeds bootstrap the resulting clusters to a category, and since each target word has a set number of senses (synset embeddings), the number of clusters per target word is pre-set. See figure 3 for a visualization. The new and unlabelled example context sentences are extracted from Korpus DK[5] and are simply word tokenized, lowercased, stripped of punctuation, considered as a BOW, and represented in the word embedding model (with the same method as decribed above for constructing synset embeddings from sense-bags). Around 1000 example sentences are extracted per targetword. We apply the K-means algorithm from the cluster package[6] included in the module Scikit-Learn (Pedregosa and Varoquaux, 2011) from Python. We set the parameter of number of clusters (*n_clusters*) to the number of synset embeddings constructed for the current target word and set the synset embeddings as initial cluster centers (*init*).

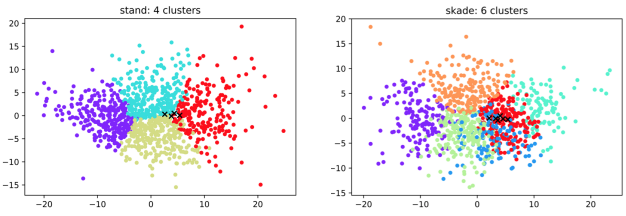

Figure 3: A 2D plot of resulting clusters of the Korpus DK example sentences for *stand* and *skade*, which have 4 and 6 synsets, and therefore 4 and 6 clusters, respectively. The black crosses are the seeds. Dimensionality reduction with PCA.

4.2. Evaluation Method

WSI systems and sense embeddings have typically been evaluated by comparing to a gold standard or in a WSD task measuring the quality by performance. In our approach, we implicitly seek to find a gold standard for word sense representations, and the quality of the developed sense representations are measured here by performance in a WSD task.

Computational semantic analysis systems are typically evaluated on the data sets from the ongoing series of SemEval, the International Workshop on Semantic Evaluation (Kilgarriff and Palmer, 2000). The evaluation data produced for SemEval 2013 task 13: *Word Sense Induction for Graded and Non-graded Senses3* is the standard data used to test WSI systems and sense embeddings. Our evaluation data, SemDaX, contains unranked sense annotations, and annotators were asked to assign one sense to the given instance.

Three test sentences from SemDaX for the Danish target-word *model* (approximately similar concept as in English) are shown below.

- *Og så havde vi kursister den luksus også at have fire fantastiske modeller at arbejde med*[7]
- *Men sådan er prisklassen konkret, og de fleste modeller bliver ofte kun produceret i et meget lille antal*[8]
- *Jeg bryder mig ikke om ordet model*[9]

It has been observed in the SemDaX corpus that almost all discrepancies among annotators were due to underspecified examples, i.e., examples where the precise word sense could not be deduced from the isolated corpus excerpt alone (Pedersen et al., 2018a). In order to account for this fact, all diverging annotations in the data set are considered to be correct (and unranked). The systems applied do not detect groups of relevant senses; they merely rank by similarity and pick the most similar sense. Since the annotated data do not contain ranked senses, and our word sense representation system does not choose a set (or cluster) of relevant senses, a direct comparison with the systems developed for SemEval 2013 task 13 with the same measures is not straightforward.

As indicated above, there might be multiple (correct) classes per instance. The combination of classes might change at every instance. We make use of an accuracy score that counts a "miss" for each instance where the system fails to identify any human-labelled sense, and a "hit" whenever it guesses at least one that matches a human label[10]. It should be noted that the system has "an advantage" in cases where annotators disagree (since more that one value is considered correct) so the results need to be analyzed together with the inter-annotator agreement. This measure is equally generous to the baselines as it is to the systems we tested.

WSD is done by maximizing the cosine similarity between the synset embeddings and the given test sentence represented as a context vector within the word embedding model. The test sentence context vector is the mean vector of the sentence considered as a bag-of-word vectors. The highest-similarity sense representation is chosen.

We apply three baselines:

- **Extended Lesk (E-lesk)**: WSD by cosine similarity between the centroid of the BOW from the wordnet definition of the word sense, and the evaluation text instance vector. (Banerjee and Pedersen, 2002)
- **Random**: WSD by chance

[5] A clustering of the annotated sentences in the evaluation data, SemDax, would be more precise, but would not be a scalable approach relying on as little annotated data as possible.

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

[7] *and we as participants then had the luxury of having four fantastic models to work with*

[8] *But that is how the price level actually is, and most models are produced in a very limited amount*

[9] *I do not like the word model*

[10] We tested the Kullback-Leibler divergence score as an alternative "soft" evaluation measure to incorporate the fact that there can be multiple correct answers, but the human distributions are far more "spiky" than the normalized system scores, leading to statistically insignificant differences between systems.

Target words	**Synsets**	**Annotated senses** (incl. idiomatic expressions)
Ansigt *(face)*	6	16
Blik *(look, glace, tin)*	6	8
Hold *(team, side, gang)*	8	10
Hul *(hole, gap, leek)*	13	22
Kort *(card, map, plan)*	10	21
Lys *(light, candle, lamp, glare)*	16	30
Model *(model, pattern, type, design)*	8	9
Plade *(plate, sheet, disc)*	13	13
Plads *(room, space, square, post)*	10	21
Skade *(harm, injury, damage, magpie, ray)*	6	12
Slag *(battle, stroke, cape, roll)*	15	28
Stand *(state, condition, shape, booth, stand)*	4	11
Stykke *(piece, part, length, paragraph)*	16	22
Top *(top, peak, apex)*	5	12
Vold *(violence, bank)*	7	10
Kontakt *(contact, switch, touch)*	7	9
Selskab *(company, party, association)*	9	11

Table 1: Target words with number of DanNet synsets (column 1) and number of senses actually encountered in the data (column 2). Some senses encountered in the annotated data are merged and link into the same synset, the reason for which we see the difference in numbers across columns.

- **Most frequent sense (MF)**

The MF as default is usually a very hard baseline to beat, in particular for the most polysemous part of the vocabulary, as we are doing here. See discussion of this in section 6.

4.3. Materials

DanNet: The Danish wordnet, DanNet, was compiled semi-automatically from the Danish dictionary *Den Danske Ordbog* (Hjorth and Kristensen, 2005). These two resources are therefore highly related and possible to link. The 65,000 synsets in DanNet are interrelated via 325,000 semantic relations. All synsets are assigned an ontological type, a corresponding supersense, and come with a definition and an example sentence. The DanNet information extracted are word collections: either words from relevant synsets (i.e., related concepts), or words from the synset example sentences and definition sentence considered as a BOW. The BOW (i.e. the synset sense-bag) is unified and represented as a centroid in the word embedding model according to the method described in 4.1.

Evaluation data: As previously mentioned, the words of interest in our work are 17 of the most polysemous Danish nouns. These words were handpicked by language experts for lexical sample studies as they are both extremely polysemous, yet frequent. See Table 1. The SemDaX corpus is a subpart of the 45 million words CLARIN Reference Corpus (Asmussen, 2012) and consists of different text types. We extract from SemDaX the 6,012 sentences containing our polysemous target nouns. These are annotated with dictionary senses by 2-6 annotators (advanced students and researchers). There are 355 sentences per target noun on average, and the more polysemous a word, the more sentences are included. For the WSD task, we include the window of 5 context words around the target noun in each annotated sentence. The text is simply lowercased and punctuation is removed. As mentioned above, every test sentence is considered as a BOW and represented as a centroid in the word embedding model (similarly as the synset sense-bags).

Note, the nouns are highly ambiguous, so a Krippendorf's α agreement of 0.80 is hard to reach here. The work of Pedersen et al. (2018) finds an agreement of 0.67 useful, which is mostly met in the agreement statistics. For relatively fine-grained sense inventories, a lower agreement score is acceptable.

The word embedding model is created by the Society for Danish Language and Literature (Sørensen and Nimb, 2018). They used the Gensim package (Řehuřek and Sojka, 2010) to train a Word2Vec model (Mikolov et al., 2013) on a corpus of roughly 920 million running words. The corpus had 6.3 million token types, where 5 million occurred less than 5 times. The dimensions of the CBOW word embeddings are 500, a window size of 5, and a threshold for rare words at 5.

Korpus DK: is a corpus [11] of different text types in Danish, and has a size of 56 million words. It consists of relatively recent language and mostly every-day language use. For each target noun, around 1000 sentences containing that noun are extracted. A window of 5 words and no normalization is chosen in line with the pre-processing of other data in this project. Every sentence is considered as a BOW and represented as a centroid in the vector space.

Software packages: With Python (van Rossum, 1995) we used the Sci-kit Learn package (Pedregosa and Varoquaux, 2011), the NLTK package (Bird et al., 2009) and SciPy (Jones et al., 2001).

Data mapping: As mentioned in the introduction, a key from dictionary senses in the evaluation data to DanNet was manually created. For 17 target nouns with 19.1 dic-

[11] `https://ordnet.dk/korpusdk`

tionary senses on average, where 15.6 senses on average was apparent in the annotated data, 159 links are found, with an average on 9.4 senses per word. See Table 1 for an overview across target words. The number of DanNet senses is slightly smaller than that of the dictionary. This is for the most part due to the many idiomatic expressions in the dictionary which are not (as they normally are not) included in the wordnet. To avoid leaving these instances out, the dictionary labels of the target noun in the figurative expressions are merged with the synset that corresponds to the literal sense of the noun. This follows the principle of annotation of idiomatic expressions (without a dictionary entry) or other figurative speech in the work of Pedersen et al. (2018a) where the annotation process is described.

5. Results

The results for all experiments are shown in Table 2. Except for the cluster centroid experiment, the results show steady improvements from .21 to .34 and exceed the random and E-lesk baseline at .13 and .16, respectively. However, the performance does not reach the MF sense baseline at .56 (discussed in Section 6.).

Sense representation	**Acc.**	Acc. ex. MF
1. Synset members	.21	.28
2. Example sentence	.26	.29
3. Example sentence+	.29	.31
4. Combination	**.34**	.36
5. Cluster centroid	.19	.22
Random	.13	.15
E-lesk	.16	.23
MF	.56	-

Table 2: WSD results

When excluding the MF class in the data and the corresponding synset embedding , the experiments actually perform slightly better and show the same steady improvements (again, except for exp. 5). Interestingly, when working with less frequent senses, the performance of exp. 1 seems to be the most improved.

6. Discussion

The best results for WSD with cosine similarity are achieved when combining all components (exp. 4): hypernyms, hyponyms, synonyms, near-synonyms, used-for, made-by semantic relations together with BOW word sense definition, the BOW example sentence, as well as and the BOW example sentences from hypo- and hypernym synsets. The more features used, the better the performance.

Synset richness: The size of and shared proportion of information of the synsets seems to be important for the sense representations in experiment 4, where the example sentence information for experiment 2-3 works best for homonyms. Experiment 4 performs worse than experiment 2, in particular in the case of the words *hold*, and *vold*, but also for *slag*, *stand*, *kontakt*, and *selskab*. Investigation of the synset member size for *hold* shows that almost half of the synsets only have one concept associated with it in DanNet, namely one hypernym. This is rather little information for establishing a synset embedding, and further, hypernyms tend to be more general and thus less informative.

Level of polysemy: Annotators report that the sentences often lack context and that the senses are highly polysemous (Pedersen et al., 2018a). Worst results from the system are found for *lys* which has a high number of senses (16), but no huge evaluation advantage since the inter-annotator agreement is relatively high (.81). Also, though the sense number is high, the senses are related in meaning and the differences are often very subtle. The target nouns *lys* and *kort* both share word form with common adjectives in Danish, which possibly affects the word embeddings. This could explain why the system performs worse for these words. The words that generally are disambiguated most satisfactorily are *blik*, *hold*, *stand*, *top* and *selskab*. All of these words have low overlap in the DanNet synsets, are homonyms, or have non-subtle sense differences.

For the word *top* and especially *stand*, the performance of experiment 4 is higher than for the other words. This might be due to the low number of senses of these words: *stand* has 4 senses, and *top* has 5, where the average number of senses is 9.4. Also, *stand* is often annotated with the same sense (and high inter-coder agreement) which suggests that there is one highly dominant sense.

In experiment 1-4, the WSD of *blik* also works well compared to the other words considering the performance of the most frequent sense. This word has a relatively low inter-annotator agreement and "only" 6 senses, which could be an explanation. This word is also a case of homonymy (i.e., unrelated meanings) which is foreseen to increase the distance between the sense embeddings in the word embedding model.

Idiomatic expressions: These expressions are relatively static in appearance. A BOW of an idiomatic expression as a sense representation vector will most likely disambiguate a corresponding context vector correctly. (See discussion of *face* below.) Now they are merged with the literal sense used in the expression, which creates bias and imprecise mapping between dictionary senses and DanNet synsets.

Clusters: Experiment 5 was motivated by the hypothesis that the best synset embeddings from former experiments might work as seeds for the clustering of more example sentence data, where the cluster centroids could function as a new synset embedding. However, the results prove otherwise, suggesting that the construction of the synset embeddings does not have clear enough information as a base for clustering.

A qualitative investigation of the sentences in the clusters confirms the results. There are patterns that begin to emerge. The target word *ansigt (face)* has 6 senses. The non-literal senses were captured in the least satisfactory way: the clusters for *face as a manifestation/appearance of a thing or phenomenon*, and *face as the character/nature of a person* contained many instances of the literal and simplest sense of *face*. The clusters of this literal sense proved to be the best and had fewer non-literal senses, although they still contained several errors. This sense was often

mixed with *face as an expression/state of mood*, which actually can also be hard for annotators to distinguish between. The cluster of *face* as a *face-like front of an object* contains mostly non-literal senses: the DanNet synset only contains *form* (same as in English) as the related lexeme and no words about persons or physiological words. This cluster contains mostly sentences about God and the Bible, which could be because the clustering algorithm followed that gradient. Finally, *face as a public profile/known face* performs relatively well and captures most instances where *kendt ansigt (known face)* and *ansigt udadtil (public/outward face)* appears in the sentence.

MF sense is hard to beat: As mentioned previously, beating a majority classifier is in general very difficult, and even more difficult when dealing with a lower-resourced language such as Danish. Our experiments indeed confirm this; however, it should be emphasized that we examine the performance of the approach when tested on the hardest task available: the most polysemous nouns in Danish. In other words, our model is expected to perform considerably better on closer-to-average polysemy words.

7. Conclusion

This study set out to determine the possibility of building appropriate sense representations for Danish by combining word embeddings with synset information from the Danish wordnet. The rationale is to combine corpus evidence with senses outlined by humans. We represented the data in a word embeddings space and tested the process in a very hard WSD task. Thousands of example sentences were auto-tagged by sense clustering,

As expected, wordnet-associated data proves to be quite informative for the WSD task. Generally speaking, the more semantic relations and information included from the wordnet, the better the results. However, the word sense representation system has room for improvement, in that the most-frequent baseline is not yet overcome in these unbalanced datasets.

Nevertheless, our sense representation system produces promising results. The best synset embeddings in our study are able to disambiguate well above chance, considering the highly polysemous selection of test words in mind (almost 20 senses on average). We expect performance to increase when handling Danish vocabulary items with closer-to-average polysemy.

For future work, we plan to enrich the synset information with data from The Danish Thesaurus, and we foresee that these enriched data could potentially improve our model. Additionally, the technique of Nieto-Piña and Johansson (2018), linking word sense embedding models to lexical resources, is interesting and could be relevant for future improvements.

Finally, it would be interesting in future to experiment with the granularity level of senses, with the exclusion of idiomatic expressions from the WSD task, and with using our sense-based word clusters to create new evaluation materials.

8. Acknowledgements

This work is partly funded by the H2020 Infraria project ELEXIS and by the Swedish Research Council project 2014-39 that funds the Center for Linguistics Theory and Studies in Probability (CLASP).

9. Bibliographical References

Agirre, E. and Edmonds, P. (2006). Word Sense Disambiguation: Algorithms and Applications. *Text Speech and Language Technology*, 33(33):384.

Asmussen, J. (2012). CLARIN-Referencekorpus. *Sprogteknologisk Workshop October 31*.

Banerjee, S. and Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Computational Linguistics and Intelligent Text Processing: Third Initernational Conference, CiCLing*, pages 136–145, Mexico City, Mexico.

Bartunov, S., Kondrashkin, D., Osokin, A., and Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Cadiz, Spain.

Bhingardive, S., Singh, D., and Murthy, R. (2015). Unsupervised Most Frequent Sense Detection using Word Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1238–1243.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language ToolKit (NLTK) Book*. O'Reilly Media Inc.

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

Chen, X., Liu, Z., and Sun, M. (2014). A Unified Model for Word Sense Representation and Disambiguation. In *Proceedings of EMNLP*, pages 1025–1035, Doha, Qatar.

Edmonds , P. and Kilgarriff, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8:279 – 291, 12.

Faralli, S., Panchenko, A., Biemann, C., and Ponzetto, S. P. (2016). Linked Disambiguated Distributional Semantic Networks. In *The 15th International Semantic Web Conference (ISWC)*, pages 56–64, Kobe, Japan.

Hjorth, E. and Kristensen, K. (2005). *Den Danske Ordbog*. Gyldendal, Copenhagen, 1 edition.

Johansson, R. and Nieto-Piña, L. (2015). Embedding a Semantic Network in a Word Space. *Naacl-2015*.

Jones, E., Oliphant, T., Peterson, P., and others. (2001). SciPy: Open source scientific tools for Python.

Kilgarriff, A. and Palmer, M. (2000). Introduction to the special issue on SENSEVAL. *Computers and the Humanities*.

Kilgarriff, A. (1997). "I don't believe in word senses". *Computers and the Humanities*.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, March.

Mancini, M., Camacho-Collados, J., Iacobacci, I., and Navigli, R. (2017). Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, Vancouver, Canada. Association for Computational Linguistics.

Martinez Alonso, H., Johannsen, A., Olsen, S., Nimb, S., Sørensen, N., Braasch, A., Søgaard, A., and Pedersen, B. (2015). Supersense tagging for danish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015*, volume 109. Linköping University Electronic Press. Der er ikke overensstemmelse mellem det ISSN-nr der står på proceedings og det der findes i databasen.

Mihalcea, R., Chklovski, T., and Kilgarriff, A. (2004). The senseval-3 english lexical sample task. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 01.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. pages 1–12.

Navigli, R. and Di Marco, A. (2013). Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. *Computational Linguistics*, 39(3):709–754.

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar. Association for Computational Linguistics.

Nieto-Piña, L. and Johansson, R. (2018). Automatically Linking Lexical Resources with Word Sense Embedding Models. In *Proceedings of SemDeep-3, the 3rd Workshop on Semantic Deep Learning*, pages 23–29, Santa Fe, New Mexico, USA.

Palmer, M., Babko-Malaya, O., and Dang, H. (2004). Different sense granularities for different applications. In *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems*, Boston, MA. HTL/NAACL.

Panchenko, A., Lopukhina, A., Ustalov, D., Lopukhin, K., Arefyev, N., Leontyev, A., and Loukachevitch, N. (2018). Russe'2018: A shared task on word sense induction for the russian language. 03.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). Dannet: The challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

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

Pedersen, B. S., Aguirrezabal Zabaleta, M., Nimb, S., Olsen, S., and Rørmann, I. (2018a). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In *Proceedings of Global WordNet Conference 2018*, pages 1–8, Singapore. Global WordNet Association.

Pedersen, B., Nimb, S., Søgaard, A., Hartmann, M., and Olsen, S. (2018b). A danish framenet lexicon and an annotated corpus used for training and evaluating a semantic frame classifier. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference, Miyazaki, Japan.* European Language Resources Association.

Pedregosa, F. and Varoquaux, G. (2011). *Scikit-learn: Machine learning in Python*.

Pelevina, M., Arefyev, N., Biemann, C., and Panchenko, A. (2017). Making Sense of Word Embeddings. (2012).

Pilehvar, M. T. and Collier, N. (2016). De-conflated semantic representations. In *Proceedings of EMNLP*, Austin, Texas.

Pradhan, S. S., Loper, E., Dligach, D., and Palmer, M. (2007). Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 87–92, Stroudsburg, PA, USA. Association for Computational Linguistics.

Řehuřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 5. ELRA.

Rothe, S. and Schütze, H. (2017). Autoextend: Combining Word Embeddings with Semantic Resources. *Computational Linguistics*, 43:3:593–617.

Saedi, C., Branco, A., António Rodrigues, J., and Silva, J. (2018). WordNet Embeddings. In *Proceedings of the 3rd Workshop on Representation Learning for NLP*, pages 122–131, Melbourne, Australia. Association for Computational Linguistics.

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

Ustalov, D., Teslenko, D., Panchenko, A., Chernoskutov, M., Biemann, C., and Ponzetto, S. P. (2018). An Unsupervised Word Sense Disambiguation System for Under-Resourced Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, 4.

van Rossum, G. (1995). Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 5.