Multilingual Modal Sense Classification using a Convolutional Neural Network

Ana Marasović and **Anette Frank**
Research Training Group AIPHES
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{marasovic,frank}@cl.uni-heidelberg.de

Abstract

Modal sense classification (MSC) is a special WSD task that depends on the meaning of the proposition in the modal's scope. We explore a CNN architecture for classifying modal sense in English and German. We show that CNNs are superior to manually designed feature-based classifiers and a standard NN classifier. We analyze the feature maps learned by the CNN and identify known and previously unattested linguistic features. We benchmark the CNN on a standard WSD task, where it compares favorably to models using sense-disambiguated target vectors.

1 Introduction

Factuality recognition (de Marneffe et al., 2012; Lee et al., 2015) is a subtask in information extraction that differentiates facts from hypotheses and speculation, expressed through signals of modality, most prominently, modal verbs and adverbs. Modal verbs are, however, ambiguous between an *epistemic* sense (possibility) as opposed to non-epistemic *deontic* (permission/obligation) or *dynamic* (capability) senses, as in: *He could be at home* (epistemic), *You can enter now* (deontic) and *Only John can solve this problem* (capability).

Modal sense classification (MSC) is a special case of sense disambiguation that is also relevant in areas of dialogue act and plan recognition in AI, as well as novel tasks such as argumentation mining. Prior work (Ruppenhofer and Rehbein, 2012; Zhou et al., 2015) addressed the task with feature-based classification. However, even with carefully designed semantic features the models have difficulties beating the majority sense baseline in cases of difficult sense distinctions and when applying the models to heterogenous text genres.

We cast modal sense classification as a novel semantic sentence classification task using a convolutional neural network (CNN) architecture. Our contributions are: (i) our experiments on MSC confirm the adequacy of CNNs for modeling propositions in *semantic* sentence classification tasks (cf. Kim (2014)); (ii) we show that automatically learned features in a CNN outperform manually designed features for difficult modal verbs and novel genres; (iii) we demonstrate that the CNN approach can be generalized across languages, by adapting the model to German. (iv) We offer insights into the linguistic properties captured by the learned feature maps. Finally, (v) we benchmark the CNN on a standard WSD task, comparing it to a WSD model using rich sense-disambiguated embeddings and obtain comparable results.

2 Prior and related work

Modal sense classification (MSC). We focus on disambiguation of modal verbs, adopting the sense inventory established in formal semantics: *epistemic, deontic/bouletic* and *circumstantial/dynamic*.[1] We compare to prior work in Ruppenhofer and Rehbein (2012) and follow-up work in Zhou et al. (2015) (henceforth, R&R and Z+). R&R induced modal sense classifiers from manual annotations on the MPQA corpus (Wiebe et al., 2005) using word-based and syntactic features. Z+ propose an extended semantically informed model that significantly outperforms R&R's results. Z+ also create heuristically sense-annotated training data from parallel corpora, to overcome sparsity and bias in the MPQA corpus. However, their models do not beat the majority sense baseline for the difficult modal verbs, *may, can* and *could*.

[1]These senses correspond to (Baker et al., 2010)'s modal categories (with *deontic* split into *requirement* and *permissive*), and R&Rs inventory, with regrouping of *concessive, conditional* and *circumstantial*, cf. Zhou et al. (2015).

Proceedings of the 1st Workshop on Representation Learning for NLP, pages 111–120,
Berlin, Germany, August 11th, 2016.

Modal sense classification interacts with genre and domain differences. Prabhakaran et al. (2012) observe strong cross-genre effects and missing generalization capacities when applying their modality classifier to out-of-domain genres.

Word Embeddings and Sense Disambiguation. Taghipour and Ng (2015) investigate the impact of word embeddings on classical WSD, using pre-trained embeddings and tuning them to the task using a NN. Both variants, integrated into the state-of-the-art system IMS (Zhong and Ng, 2012), improve WSD performance on benchmark tasks.

Ordinary word embeddings do not differentiate word senses. Rothe and Schütze (2015) explore supervised WSD using sense-specific embeddings, which they induce by exploiting *sense encodings* and *constraints* given by a lexical resource.[2] Integrating the sense-specific vectors into IMS yields significant improvements and small gains relative to Taghipour and Ng (2015). Hence, word embeddings – tuned to the task or sense-specific – prove beneficial for supervised WSD.

The CNN approach we investigate in our work does not employ a fixed feature space or a predefined window around the target word. It flexibly learns feature maps for variable window sizes over the embedding matrix for the full sentence. In contrast to Rothe and Schütze (2015), embeddings used by our CNNs models are knowledge-lean and do not encode senses of the target words.

Sentence classification using CNNs. Recent work investigates NN architectures and their ability to capture the semantics of sentences for various classification tasks. Kalchbrenner et al. (2014) construct a dynamic CNN that builds on unparsed input and achieves performance beyond strong baselines for sentiment and question type classification. By contrast, recursive neural networks (Socher et al., 2013) take parsed input, recursively generate representations for intermediate phrases, and perform classification on the basis of the full sentence representation.

Kim (2014) evaluates a one-layer CNN on various benchmark tasks for sentence classification. CNNs trained on pre-trained (static) embeddings perform well and can be further improved by tuning them to the task (non-static). Using two channels did not significantly improve results. Overall, the CNNs show consistently strong performance, improving on state-of-the-art results in 4 out of 7 tasks, i.a., sentiment and opinion classification.

3 A CNN for modal sense classification

We aim at a NN approach to MSC that (i) improves over existing feature-based classifiers, (ii) alleviates manual crafting of features, (iii) generalizes over various text genres, and (iv) is easily portable to novel languages. Besides this, MSC is a special kind of WSD, in that modal verbs have a restricted sense inventory shared across languages, and act as operators that take a full proposition as argument. We thus cast MSC as a semantic sentence classification task in a CNN architecture, adopting the one-layer CNN model of Kim (2014), a variant of Collobert et al. (2011). Unlike Kim (2014) we will use only one channel, but experiment with various types of word vectors.

A CNN represents a sentence with a fixed size vector, passed to classifier to classify the sentence into task-specific target categories. In our case, it will classify sentences into three modal sense categories. The input layer is a matrix $\mathbf{x} \in \mathbb{R}^{s \times d}$, with each row corresponding to a d-dimensional word embedding $x_i \in \mathbb{R}^d$ of a word in the sentence of length s. Word embeddings can be randomly initialized or pre-trained vectors, e.g. word2vec (Mikolov et al., 2013) or dependency-based (Levy and Goldberg, 2014) embeddings. Based on the input layer, a CNN builds up one or more convolutional layers. A convolution is an operation between sub-matrices of the input matrix $\mathbf{x} \in \mathbb{R}^{s \times d}$ and a *filter* parametrised by a weight matrix $\mathbf{w} \in \mathbb{R}^{n \times d}$, that returns a vector usually referred to as a *feature map*. Formally, let $\mathbf{x}_{i-n+1:i}$ be the sub-matrix of the input matrix $\mathbf{x}$ from the $(i-n+1)$-th row to the i-th row and let $\langle .\,, . \rangle_{\mathcal{F}}$ denote the sum of elements of the component-wise inner product of two matrices, known as Frobenius inner product. The i-th component of the feature map $\mathbf{c}$ is obtained by taking the Frobenius inner product of the sub-matrix $\mathbf{x}_{i-n+1:i}$ with the filter matrix $\mathbf{w}$

$$c_i = \langle \mathbf{x}_{i-n+1:i}, \mathbf{w} \rangle_{\mathcal{F}} \,, \tag{1}$$

for $i \in \{n, \ldots, s\}^3$. Afterwards, we add a bias term, $b \in \mathbb{R}$ to every component of the feature map and apply an activation function f,

$$\tilde{c}_i = f(c_i + b) \,. \tag{2}$$

[2] Modal verbs are not or not systematically covered in WordNets or VerbNet; FrameNet relates modal verbs to their predominant sense only. Also, FrameNet's frame-to-frame relations are known to lack coverage (Burchardt et al., 2009).

[3] We apply the narrow type of convolution.

Finally, *max-over-time pooling* (Collobert et al., 2011) is applied over a single feature map that extracts the maximum value $\hat{c} = max\{\tilde{\mathbf{c}}\}$, which represents the chosen feature for this feature map. Like Kim (2014) we don't use just one filter as described, but multiple filters with different region sizes n, resulting in multiple feature maps. Features obtained through max-pooling from each feature map are concatenated to a vector representation of the input sentence that is passed to the softmax layer. Parameters to learn are elements of the filter matrices and the input matrix when word vectors are tuned.

Filters are trained to be especially active when they encounter a sequence of words relevant for the given classification task. Kalchbrenner et al. (2014) present n-grams of different feature detectors that capture positive or negative sentiment phrases, and also more abstract semantic categories, such as negation or degree particles ('too') that are relevant in compositional sentiment detection. In the modal sense classification task, we expect the feature maps to capture semantic categories found to be relevant in prior work, such as tense, aspectual classes, negation and semantic properties of verbs and phrases. Moreover, prior work has shown that MSC profits from features that model the wider syntactic context, esp. subject and embedded verb and their semantics (abstractness, semantic class, aspect, tense). Explicit modeling of these features as in Z+ improves performance, but requires feature design for each new language. Also, modeling semantic features through lexical resources is subject to sparsity, and relying on parsed input leads to lack of robustness.

Given that MSC profits from semantic features in the wider syntactic context, we expect that a CNN that applies filters of variable sizes to various regions of the sentence to learn feature maps can capture diverse linguistic features, and offers greater flexibility compared to a conventional WSD model with a fixed window size centered around the target word. To investigate these special properties of the CNN model, we test it on English and German data. While in English, subject, modal and embedded verb are in a close syntactic context, in German, they can be distributed over wider distances, and the feature maps are expected to capture properties over wider distances.

We perform experiments for MSC for English and German, using various data sets. Section 4 presents the data, experimental settings and the model variations we investigate. We perform detailed quantitative and qualitative evaluation of our experimental results. In Section 5, we evaluate the CNN approach in a lexical sample WSD task, to benchmark its performance on a well-studied data set, and to investigate the potential advantage of learning feature maps based on flexible window sizes. To our knowledge, this constitutes the first attempt to apply a CNN model in a WSD task.

4 Modal sense classification

4.1 Data

Our experiments are based on three data sets. Their basic composition is given in Table 1.[4]

1) MPQA + EPOS$_E$ The English benchmark data set MPQA from R&R was further enriched through balanced heuristically tagged training data, EPOS$_E$, by Z+. The EPOS$_E$ data set was obtained using a cross-lingual sense projection approach. Z+ identified paraphrases for modal senses (e.g. *brauchen-need; erlauben-permit* for deontic, *schaffen-able to* for dynamic sense), extracted sentences from a parallel corpus with a modal verb aligned to a sense-identifying paraphrase, and tagged them with the identified modal sense. Z+ measured 0.92 accuracy on 420 instances of the heuristically tagged corpora. To alleviate distributional bias stemming from the MPQA dataset, Z+ balanced the blend of MPQA with EPOS$_E$ using under- and oversampling. We experiment with both versions ($\pm$ balanced).[5]

2) MASC A subset of the multi-genre corpus MASC (Ide et al., 2008), consisting of 19 genres was manually annotated (Anonymous) with modal senses for the same modal verbs. The annotated data consists of $\approx$100 instances for each genre.[6]

3) EPOS$_G$ Following the method of Z+, we constructed a German data set EPOS$_G$ from the Europarl and OpenSubtitles corpora of OPUS (Tiedemann, 2012) by projecting modal sense categories from English to German, using selected modal sense identifying English paraphrases. The resulting corpus with sense-tagged German modal verbs

[4]More detailed information will be provided through accompanying material with the final version. The annotated MASC and EPOS$_G$ data sets will be made publicly available.

[5]Their data is publicly available through their website. We omit *shall* from MPQA, due to low number of occurrences.

[6]Exceptions with less than 100 instances are journal, newspaper, technical, travel guides, and telephone.

		can	could	may	must	should
MPQA	ep	2	156	130	11	26
	de	115	17	9	83	248
	dy	271	67	–	–	–
EPOS$_E$	ep	150	40	950	800	150
	de	150	40	950	800	150
	dy	150	40	–	–	–
MASC	ep	88	144	217	29	27
	de	72	16	43	115	224
	dy	710	251	3	–	–

		dürfen	können	müssen	sollen
EPOS$_G$ (train)	ep	1000	1000	1000	1000
	de	1000	1000	1000	1000
	dy	–	1000	–	–
EPOS$_G$ (test)	ep	98	100	32	100
	de	98	47	100	100
	dy	–	100	–	–

Table 1: Composition of MPQA, EPOS$_E$, MASC and EPOS$_G$

können (can), müssen (must), sollen (should), dürfen (may) consists of a manually validated test section consisting of up to 100 instances for each sense. Annotation was done by two independent judges and one adjudicator. Balanced training data of 1000 instances per sense for each modal verb was constructed from heuristically tagged sentences that were judged high-quality by validating 20 instances for each paraphrase. For modal verbs with rare extractions, we added training data from modal verbs of shared senses, changing their verb forms to the verb form of the target verb.[7]

4.2 Experimental settings

MSC on MPQA using CNN-E$_{\mathbf{B}}$ and CNN-E$_{\mathbf{U}}$, CV For MSC we benchmark the CNN approach against the latest state-of-the-art results in Z+. We reimplemented their maximum entropy classifier (henceforth, *MaxEnt*) and trained it on their balanced and unbalanced blend of MPQA and EPOS.[8] As in Z+ we train independent classifiers for each modal verb on their respective training data.[9] For evaluation, we perform 5-fold cross validation as in Z+. Each fold for training holds a stratified 80% section of the MPQA data together with the full EPOS$_E$ data set, and we use the remaining 20% of MPQA data for testing. We refer to the CNN models trained on the $\pm$balanced versions of this data as CNN-E$_{\mathbf{B}}$ and CNN-E$_{\mathbf{U}}$.

MSC on MASC using CNN-E$_{\mathbf{B}}$ and CNN-E$_{\mathbf{U}}$ Besides MPQA, we evaluate the CNN on the multi-genre MASC (sub)corpus. For comparability with Z+, for training we use one training fold from the previous setting,[10] and evaluate on MASC as test. We analyze the performance of the CNN model overall and on different genre subcorpora (not reported here).

Both English data sets are characterized by modest training set sizes and involve a considerable distributional biases, with high most frequent sense majority baselines (cf. Tables 3 and 4).

MSC on EPOS$_G$ using CNN-G In constrast to the English data sets, the German EPOS$_G$ data set provides larger training set sizes of 1000 instances for all modal verbs and senses. This eliminates distributional bias from the data, so that the discriminating power of the classifier model is not masqued by distributional information.

4.3 Model variations

Hyperparameters Model-specific hyperparameters of the CNN are the number of filters, filter region size, and the depth of the network. We restrict our model to a one-dimensional CNN architecture.

Following the advices in Zhang and Wallace (2015), we used following setting: ReLU (rectified linear unit) as activation function, filter region sizes of 3, 4, and 5 with 100 feature maps each, dropout keep probability of 0.5, l_2 regularisation coefficient of 10^{-3}, number of iterations of 1001[11] and mini-batch size of 50. Training is done with the Adam optimisation algorithm (Kingma and Ba, 2014) with learning rate of 10^{-4}. Filter weights are initialized using Glorot-Bengio strategy (Glorot and Bengio, 2010). We experimented with some parameter variations (using nested CV), but found no consistently better results. In all following MSC experiments we thus used this hyperparameter setting for CNN training.

Word embeddings In the first and third experimental setting we investigate the impact of static and tuned versions of different word vectors: word2vec (Mikolov et al., 2013), dependency-based (Levy and Goldberg, 2014) and randomly initialized embeddings.

We used publicly available `word2vec` vectors that were trained on Google News for En-

[7] Replacing e.g. *könnte* with *dürfte* in *Es könnte Dir gefallen* extracted from *You might get a taste for it.*

[8] We omit *shall* with a small number of instances.

[9] This holds for all our experiments.

[10] Hence, one 80% fold of MPQA plus EPOS$_E$. Despite this small difference, we refer to the CNN models as above, as CNN-E$_{\mathbf{B}}$ and CNN-E$_{\mathbf{U}}$.

[11] We did not perform early stopping.

glish[12] and various datasets for German (Reimers et al., 2014)[13], as well as English dependency-based vectors trained on Wikipedia[14]. The German dependency-based embeddings were trained on the SdeWaC corpus (Faaß and Eckart, 2013), parsed with Malt parser. We used 300 dimensions for English embeddings and 100 for German.

For words without a pre-trained vector and in the random initialization setting, each dimension of the random vector was sampled from $\mathcal{U} \sim [-a, a]$ with parameter a picked such that the variance of the uniform distribution equals the variance of the available pre-trained vectors.

Baselines For MPQA and MASC, the classifiers are compared against strong *majority sense baselines, BL_{maj}*, due to skewed sense distributions in the training data. Further, we compare the CNN results to the reconstructed *MaxEnt* classifier from Z+, trained on the blend of MPQA and EPOS with R&R's shallow lexical and syntactic path features and the newly designed semantic features of Z+.

To our knowledge, there is no work on modal sense classification using a neural network. We thus compare our CNN models with a simple, one-layer neural network NN to investigate the impact offered by the more complex CNN architecture.

Input to the NN is the sum of all vectors of the words in the sentence. As for the CNN, we experimented with different types of word vectors.

The *hyperparameter setting* for the NN is: ReLU as activation function, l_2 regularisation coefficient of 10^{-3}, hidden layer size of 1024, number of iterations of 3001, dropout keep probability of 0.5, and mini-batch size of 50. Training is again done with the Adam optimisation algorithm (Kingma and Ba, 2014) with learning rate of 10^{-4}. Weights are initialized using Glorot-Bengio strategy (Glorot and Bengio, 2010).[15]

4.4 Results

English

In Table 2 we report results for CNN-$\mathrm{E_B}$ and CNN-$\mathrm{E_U}$ with diverse input representations. For balanced training, dependency based vectors yield the best (*can*, *could*) or equally good results (*may*, *must*, *should*). *Could* is the only case with large performance differences depending on the choice of embeddings. For *can* and *could* choosing either static or tuned versions of vectors is beneficial. With unbalanced training, dependency-based vectors are outperformed by `word2vec` for *must* and by randomly initialized vectors for *can*. Large differences in the results for *could* w.r.t. the choice of embeddings, are no longer present.

CNN-$\mathrm{E_B}$	can	could	may	must	should
w2v-static	65.02	51.67	**93.57**	**93.82**	**90.77**
w2v-tuned	63.73	54.17	**93.57**	**93.82**	**90.77**
dep-static	**65.78**	56.67	**93.57**	**93.82**	**90.77**
dep-tuned	59.89	**67.50**	**93.57**	93.29	90.42
rand-static	63.99	46.67	**93.57**	92.79	**90.77**
rand-tuned	64.50	48.33	**93.57**	92.79	**90.77**
CNN-$\mathrm{E_U}$	**can**	**could**	**may**	**must**	**should**
w2v-static	70.10	65.27	**93.49**	**94.97**	**90.59**
w2v-tuned	70.62	66.10	**93.49**	**94.97**	**90.59**
dep-static	69.85	65.27	**93.49**	94.46	**90.59**
dep-tuned	69.59	**66.55**	**93.49**	93.95	**90.59**
rand-static	70.36	64.45	**93.49**	93.45	**90.59**
rand-tuned	**70.87**	64.86	**93.49**	93.45	**90.59**

CNN-G	dürfen	können	müssen	sollen
w2v-static	91.92	68.82	77.61	71.64
w2v-tuned	**99.49**	74.09	83.58	72.14
dep-static	91.92	63.56	75.37	73.13
dep-tuned	97.47	73.28	82.83	**74.63**
rand-static	96.46	77.33	81.34	74.13
rand-tuned	98.48	**78.95**	**85.07**	73.63

Table 2: CV accuracy for CNN-$\mathrm{E_B}$, CNN-$\mathrm{E_U}$, test accuracy for CNN-G, with different input representations.

In Table 3 we report overall results for CNN-$\mathrm{E_B}$ and CNN-$\mathrm{E_U}$ on MPQA compared to the baselines. As representations for the NN and CNN we selected, for each modal verb, the embedding type that yielded the best results (Table 2)[16].

For each training data set, scores of the CNN which are significantly better[17] than the next lower score among the baselines are <u>underlined</u>. If CNN does not yield the best results, significance between the baseline with the best score and CNN is reported. $\overline{\text{Overlining}}$ is used if CNN with unbalanced training performs significantly better than CNN with balanced training, and vice versa.

With balanced training, CNN outperforms all baselines for every modal verb and in terms of micro average. However, differences between CNN

[12] https://code.google.com/archive/p/word2vec

[13] https://www.ukp.tu-darmstadt.de/research/ukp-in-challenges/germeval-2014

[14] https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings

[15] This is clearly not shown to be the best hyperparameter setting, as we chose it heuristically without tuning.

[16] For NN the impact of word vectors was investigated as well.

[17] By conducting the mid-p-value McNemar test (Fagerland et al., 2013) with p <0.05.

	can	could	may	must	should	micro
BL_{rand}	33.33	33.33	50.00	50.00	50.00	41.49
MaxEnt	59.64	61.25	92.14	87.60	90.11	74.88
NN	56.01	55.42	90.00	75.24	88.68	69.74
CNN-E_B	**65.78**	**67.50**	**93.57**	**93.82**	**90.77**	**79.29**

	can	could	may	must	should	micro
BL_{maj}	69.92	65.00	93.57	94.32	90.81	80.18
MaxEnt	64.76	63.33	92.14	92.78	**91.48**	78.01
NN	67.29	66.08	**94.23**	86.37	90.96	77.93
CNN-E_U	**70.87**	**66.55**	93.49	**94.97**	90.59	**80.74**

Table 3: Comparison of CV accuracies on MPQA of CNN-E_B (upper table) and CNN-E_U (lower table) with baselines.

	can	could	may	must	should	micro
BL_{rand}	33.52	33.82	48.67	46.87	46.01	38.63
MaxEnt	66.74	62.86	**87.83**	83.33	84.06	72.25
CNN-E_B	**80.46**	**64.48**	86.69	**84.72**	**88.84**	**79.33**

	can	could	may	must	should	micro
BL_{maj}	**81.61**	35.04	82.51	79. 86	89.24	72.86
MaxEnt	73.17	**55.34**	**87.45**	86.11	**89.64**	74.41
CNN-E_U	81.03	49.15	86.31	**86.80**	89.24	**76.49**

Table 4: Accuracies on MASC dataset of classifiers trained on MPQA+$EPOS_E$.

	dürfen	können	müssen	sollen	micro
BL_{rand}	50.00	33.33	50.00	50.00	39.10
NN	77.73	43.32	73.88	50.25	57.69
CNN-G	**99.49**	**78.95**	**85.07**	**74.63**	**84.10**

Table 5: Average accuracy on $EPOS_G$.

and *MaxEnt* are significant only for *can*, *could* and micro average. Moving to unbalanced training, CNN has difficulties beating the baselines (cf. *may*, *should*), but yields the best micro average. Unbalanced training for CNN outperforms balanced training in terms of micro averages, however the difference is not significant.

Table 4 summarizes the evaluation of CNN-E_B and CNN-E_U on the MASC corpus. Note that CNN with unbalanced training, CNN-E_U, does not have enough generalization capability when applied to different genres. This behavior coincides with changes of the predominant sense between training and test. CNN-E_U, as well as *MaxEnt*, is highly sensitive to such distributional changes. Even though balanced training for CNN leads to a slightly worse micro average when evaluated on MPQA, on MASC CNN–E_B yields a +3pp gain in micro average compared to unbalanced training.[18]

In sum, our evaluation shows that the CNN model is able to outperform strong baselines in most configurations. Balanced training shows more consistent results beyond the baselines and is competitive with unbalanced training, without significant difference except for *can*. In view of genre differences in MASC, the CNN–E_B model is more robust against sense changes, and yields overall better results. The strong behaviour on balanced training data shows that the CNN model is able to learn meaningful structure from the data.

German

In Table 2 we report results for CNN-G with diverse input representations. Reasons for the slightly weaker performance of dependency-based vectors compared to word2vec (1-2 pp.) can be seen in the smaller size of the training corpus, and possibly greater noise due to parsing errors.

In Table 5 we report overall results for CNN-G compared to the NN baseline.[19] The CNN outperforms both baselines by large margins, per modal verb and in terms of micro average. Given we employed perfectly balanced training data, the classifier performances reflect their ability to learn characteristic information for the classes. Indeed, the NN has great difficulties distinguishing the senses for *können* (3 senses) and *sollen*, and is outperformed by CNN-G by +35.6 and +24.4 pp. gains. The confusion matrices for CNN-G show a clear separation of these classes, in contrast to the NN.

While German is a more difficult language than English due to its syntactic properties (word order, degree of inflection), CNN-G reaches overall higher performance levels compared to English, especially for difficult cases.[20] One reason can be the morphological distinction between indicative and subjunctive (Konjunktiv), which – in interaction with tense and other factors – can ease the distinction of epistemic vs. deontic/dynamic sense. For *sollen* this morphological division is masqued, and this can explain the weaker results compared to other binary classes. Generally, CNN-G profits from larger and perfectly balanced training data.

[18]In contrast to *MaxEnt*, which does not profit from balanced training.

[19]We did not construct a MaxEnt classifier for German. For NN and CNN-G we chose the best performing embedding types per modal verb.

[20]Clearly, we cannot draw any strict comparison here.

4.5 Semantic feature detectors

Z+ provided a thorough analysis of the impact of semantic features by ablating individual feature groups. Their ablation analysis confirmed that feature groups relating to tense and aspect of the embedded verb, negation, abstractness of the subject and semantic features of the embedded verb yield significant effects on classification performance.

For *must*, Z+ found clear patterns for the occurrence of specific features and the ability to properly classify a specific sense. However, they did not identify precise features that differentiate epistemic and dynamic readings with *can*. We spefically investigated whether the learned filters for *must* can be related to the semantic categories Z+ found to be important for distinguishing its senses. In addition, we investigated whether the CNN is able to capture unattested features that differentiate epistemic and dynamic readings with *can*.

For every modal verb and every filter, we sort sentences in the training data by the maximum value obtained by applying *1-max pooling* to the feature map acquired by applying the respective filter to a sentence. For each filter and each of the top-ranked 15 sentences, we extract the ngram that corresponds to the maximum value w.r.t. the filter, i.e. the argmax of the feature map. The ngram vector is the sum of all vectors of words in the ngram. The obtained ngram vectors were plotted using the t-SNE algorithm (Van der Maaten and Hinton, 2008) and textually displayed with their surrounding context.

For *must* we found many feature detectors that relate to observations in Z+. Many filters detect past (***you must have been** out last night*; ep) vs. non-past (***we must make** further efforts*; de) and a dynamic event (***we must develop** a policy*; de) vs. stative (***you must think** me a perfect fool*; ep) reading of the embedded verb. Among others the feature detectors capture passive constructions (*actual steps **must be taken***; de) and negation (***we must not** fear*; de). Some filters were trained to capture domain vocabulary which intuitively goes along with deontic sense (*European parliament; present regulation; fisheries policy*). One filter captures telic clauses (*to address these problems; to prevent both forum; to exert maximum influence*), identifying deontic sense. Novel features not considered in Z+ are discourse markers (*but; and (then)*) that correlate with deontic sense. All in all, the CNN learns meaningful features that are known to be important for differentiating senses for *must*, and in contrast to manual feature design, it detects relevant unattested features by itself.

For *can* many filters recognise accomplishments which go along with dynamic sense, e.g. *You can do it/make it to NY*. Others detect words indicating *possibility* (ep), negation (de), discourse markers, animate subject (de and dy), passive construction (de and dy). However, without a systematic classification of these features it remains unclear how important they are for differentiating the senses of *can*. Also, similar to Z+ we did not find clear-cut features that recognize epistemic sense.

We performed a corresponding analysis of feature maps for German, following the same extraction procedure. We found the typical state (ep) vs. event (de) contrast for the embedded verb, negation and tense, and again previously unattested factors such as discourse relation markers[21] (*but; without; thereby; in order to* (dy)). For German we identified various indicators for epistemic sense (for *müssen* and *können*): attitude predicates (*believe, not know; tell me; have an idea, be afraid*), adverbials (*possibly*), conditionals (*if*); counterfactual and negative polarity contexts (*not be the case; how; ever*). Further detectors for epistemic sense are abstract subjects: placeholders for propositions (*it*), abstract concepts (*idea; music; grades; application*); indefinite subjects (*one*). We find a tendency for 1st or 2nd person subjects to co-occur with de/dy and 3rd person pronouns with ep. For *können* (dy) we find achievements (*present report; move mountains; find compromise*). For deontic readings, next to negation with 1st and 2nd person we find typical verb-object combinations for actions that can be granted: *use telephone; communicate with third parties*.

We extracted statistics about the distance of the extracted ngrams from the modal verb (distance overall; to the left/right and ngrams starting with the modal). There are no greater overall distances for German compared to English. However, for German we find significantly more ngrams that include the modal verb, especially for epistemic readings of *können, müssen, dürfen* that clearly mark subjunctive mood, whereas for *sollen*, with ambiguous forms for subjunctive and past tense, no such tendency is observed. Thus, the feature maps identify subjunctive marking (in conjunction with other factors) as relevant for classifying epis-

[21] For reasons of space we provide translations to English.

temic sense, whereas for *sollen* the lack of this indicator goes along with lower performance. Finally, we observe, for English and German, strikingly larger distances to the left of the modal verb for epistemic readings compared to non-epistemic readings. This can be traced back to indicators in the wider left-embedding context: embedding predicates, subjects, if clauses, etc.

5 Word sense disambiguation

Next to modal sense classification, we evaluate our CNN model in a classical WSD task. As benchmark corpus we chose the *SensEval-3 lexical sample* data set (Mihalcea et al., 2004), which was recently applied in Rothe and Schütze (2015) (henceforth R&S) and Taghipour and Ng (2015), using sense-specific embeddings and a NN architecture, respectively (cf. Section 2).

The training data size for the 57 target word types ranges from 14 to 263 instances. Sense labels of test instances of a given target word are predicted using the CNN model trained on the training instances for the respective word type.[22] We set the CNN hyperparameters to be the same as for MSC, except for mini-batch size and region sizes. Since the training data for some words is below 50 instances, mini-batch size was set to 10. For tuning of the region sizes, we split the training data for each word (80:20 for training and validation) and used static `word2vec` for the input representation. Among $\{(1,2,3),(2,3,4),(3,4,5),(4,5,6),(5,6,7)\}$ the best results were obtained for $(5,6,7)$.[23]

The final hyperparameter setting was used to investigate the impact of representations. Among `word2vec`, dependency-based and randomly initialised, `word2vec` performed the best, the tuned version being slightly better than static vectors. We report results for tuned `word2vec` vectors.

We compare our results to the results R&S obtained when using only sense-specific embeddings. These are not the state-of-the-art WSD results they obtain with additional features, namely POS tags of words in a small window around the target word, their discrete representation and local collocations. For sentence representation, R&S used every word in the target word sentence. For

[22]Training instances in the *SensEval-3* dataset can have more than one sense label. For training we randomly picked one of possible labels. Instances which contain more than one marked target word were omitted.

[23]However, the differences in the results were minor.

S_{naive}-prod	62.20	S-prod	64.30
S-cosine	60.50	S-raw	63.10
		CNN	**66.50**

Table 6: WSD accuracy on *SensEval-3* dataset.

sense prediction, they used the following feature vectors that are fed into a linear SVM classifier:

$$\text{S-cosine} = \langle \cos(c, s^{(1)}), \ldots, \cos(c, s^{(k)}) \rangle \,,$$
$$\text{S-product} = \langle c_1 s_1^{(1)}, \ldots, c_n s_n^{(1)}, \ldots, c_1 s_1^{(k)}, \ldots, c_n s_n^{(k)} \rangle \,,$$
$$\text{S-raw} = \langle c_1, \ldots, c_n, \ldots, s_1^{(k)}, \ldots, s_n^{(k)} \rangle \,,$$

where w is a target word with k senses, c is the centroid defined as the sum of all `word2vec` vectors of words in the sentence and $s^{(j)}$ is the embedding of the j-th synset of w.[24] They propose a variant of the *S-prod* feature vector, *S_{naive}-prod*, for which the synset embeddings are the sum of the `word2vec` vectors of all words in that sysnet.

The results are summarised in Table 6. The CNN model compares favorably to the competitor models of R&S using *AutoExtend* embeddings for WSD. It achieves slightly higher results without explicitly marking the target word, whereas the *AutoExtend* embeddings encode much richer information: what is the target word, how many possible sense it has, and knowledge-intense sense embeddings for each of its synsets. The CNN is able to compete with the rich *AutoExtend* model, and future work needs to investigate whether – similar to the S-product setting in R&S – the CNN model can achieve competitive state-of-the-art results by incorporating features corresponding to those of the IMS system of Zhong and Ng (2010).

6 Conclusion and future work

We presented an account for multilingual modal sense classification using a CNN architecture. We apply the same architecture in a standard WSD task and achieve competitive results compared to a system using richer embedding information.

Our one-layer CNN architecture outperforms strong baselines and prior art for MSC in English, including a NN and MaxEnt model, and proves particularly robust in cross-genre classification.

We applied the CNN model to German, on a data set of modest size, obtained using cross-lingual projection techniques. The CNN-G classifier outperforms a NN model by large margins.

[24]Obtained using the *AutoExtend* method of R&S.

Our approach can be easily generalized to novel languages without tedious and resource-intensive feature engineering. Through analysis of learned feature maps we gave evidence that the CNN learns both known and novel features for MSC.

The attractiveness of the CNN framework lies in its ability to learn (semantic) features from flexible window regions without syntactic processing, and the ensuing robustness on difficult text genres and its ease in generalizing to novel languages.

Acknowledgments

We thank Mengfei Zhou for her support with the German corpus construction. This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1.

References

Kathryn Baker, Michael Bloodgood, Bonnie J Dorr, Nathaniel W Filardo, Lori Levin, and Christine Piatko. 2010. A Modality Lexicon and its use in Automatic Tagging. In *Proceedings of LREC*, pages 1402–1407.

Aljoscha Burchardt, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Natural Langugae Engineering*, 15(4):527–550.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did It Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, 38(2):301–333. Special Issue: Modality and Negation.

Gertrud Faaß and Kerstin Eckart. 2013. *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25-27, 2013. Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Morten W Fagerland, Stian Lydersen, and Petter Laake. 2013. The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC medical research methodology*, 13(1):1.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Charles Fillmore. 2008. MASC: The manually annotated sub-corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-2008)*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665, Baltimore, Maryland.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 17461751, Doha, Qatar.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal, September.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL (2)*, pages 302–308.

R. Mihalcea, T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text [CD-ROM]*, pages 25–28.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea, July.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. Germeval-2014: Nested named entity recognition with neural networks. In Gertrud Faaß and Josef Ruppenhofer, editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 117–120, Hildesheim, October. Universitätsverlag Hildesheim.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803, Beijing, China.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can !? Annotating the senses of English modal verbs. In *Proceedings of the LREC 2012 Conference*, pages 1538–1545.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA.

Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *The 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 314–323.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC-2012*, pages 2214–2218, Istanbul, Turkey, May.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165 – 210.

Ye Zhang and Byron C. Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Technical report, University of Texas at Austin. arXiv:1510.03820v2.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics.

Zhi Zhong and Hwee Tou Ng. 2012. Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–282, Jeju Island, Korea.

Mengfei Zhou, Anette Frank, Annemarie Friedrich, and Alexis Palmer. 2015. Semantically Enriched Models for Modal Sense Classification. In *Proceedings of the EMNLP 2015 Workshop LSDSem: Linking Models of Lexical, Sentential and Discourse-level Semantics*, Lisbon, Portugal.