Modal sense classification

using a convolutional neural network

Ana Marasović

Institut für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg

01.07.2016.

Modal verbs are ambiguous between the following senses:

1. **epistemic** (possibility)

   He *could* be at home.

2. **deontic** (permission/obligation)

   You *can* enter now.

3. **dynamic** (capability)

   Only John *can* solve this problem.

MSC is special case of WSD

Mein Gott, sie _______ sich schrecklich gefühlt haben!

Why do we care about it?

Distinguishing facts from hypotheses and speculations, or apprehended, planned, desired states of affairs

- planned (positively): should, must + deontic
- apprehended (negative): should not + deontic
- disliked or forbidden (negative): may not +deontic
- desired (positive): should + deontic

Tasks of relevance

- factuality recognition
- sentiment analysis
- opinion mining
- argumentation
- opinion summarization

Related work

- Ruppenhofer and Rehbein (2012) → R&R
  - relatively high performance
  - shallow lexical and syntactic features
  - small-scale manually annotated corpora
  - large distributional bias
- Zhou et al. (2015) → Z+

Related work

- Ruppenhofer and Rehbein (2012) → R&R
- Zhou et al. (2015) → Z+

R&R $\boxed{\text{sparsity}} \xrightarrow{Z+} \boxed{\text{heuristically tagged data}}$

Related work

- Ruppenhofer and Rehbein (2012) → R&R
- Zhou et al. (2015) → Z+

R&R

- sparsity $\xrightarrow{Z+}$ heuristically tagged data
- distributional bias $\xrightarrow{Z+}$ balancing of the data

Related work

- Ruppenhofer and Rehbein (2012) → R&R
- Zhou et al. (2015) → Z+

R&R	sparsity	$\xrightarrow{\text{Z+}}$	heuristically tagged data
	distributional bias	$\xrightarrow{\text{Z+}}$	balancing of the data
	shallow lexical and syntactic features	$\xrightarrow{\text{Z+}}$	enriching with semantic features

Related work

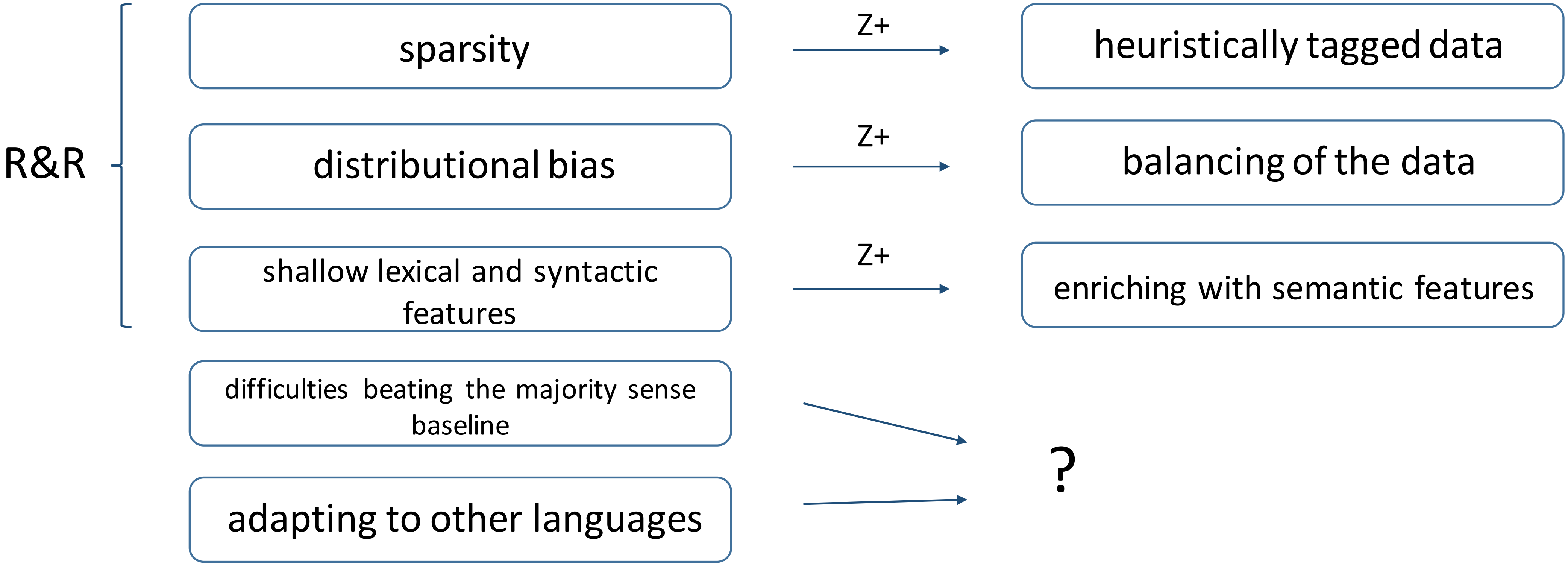
- Ruppenhofer and Rehbein (2012) → R&R
- Zhou et al. (2015) → Z+

R&R

sparsity	$\xrightarrow{\text{Z+}}$	heuristically tagged data
distributional bias	$\xrightarrow{\text{Z+}}$	balancing of the data
shallow lexical and syntactic features	$\xrightarrow{\text{Z+}}$	enriching with semantic features
difficulties beating the majority sense baseline	$\longrightarrow$	?

Related work

- Ruppenhofer and Rehbein (2012) → R&R
- Zhou et al. (2015) → Z+



Related work

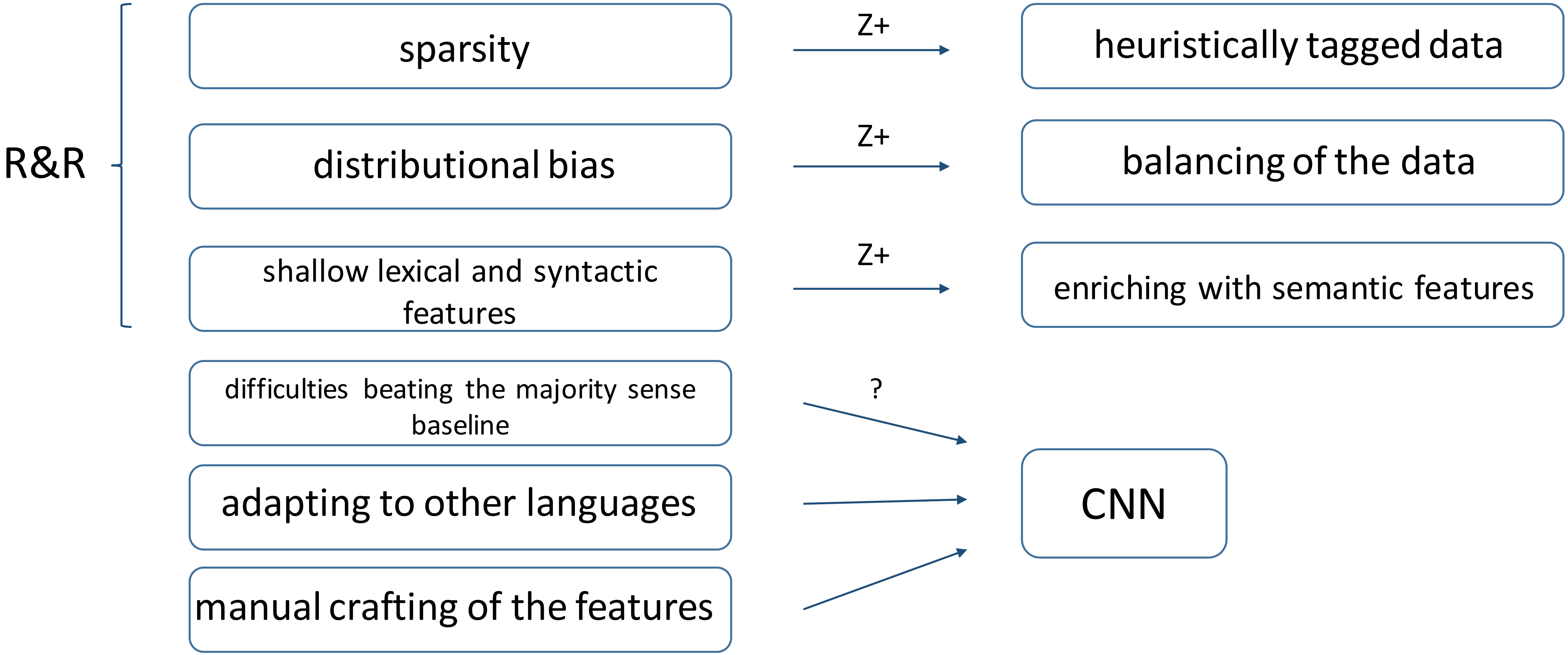
- Ruppenhofer and Rehbein (2012) → R&R
- Zhou et al. (2015) → Z+

R&R

sparsity	$\xrightarrow{\text{Z+}}$	heuristically tagged data
distributional bias	$\xrightarrow{\text{Z+}}$	balancing of the data
shallow lexical and syntactic features	$\xrightarrow{\text{Z+}}$	enriching with semantic features
difficulties beating the majority sense baseline	→	?
adapting to other languages	→	?
manual crafting of the features	→	?

Related work

- Ruppenhofer and Rehbein (2012) → R&R
- Zhou et al. (2015) → Z+



Outline

- Introduction
- Convolutional neural network (CNN) for sentence modeling
- CNN for MSC
- CNN for general word sense disambiguation (WSD)
- Future work

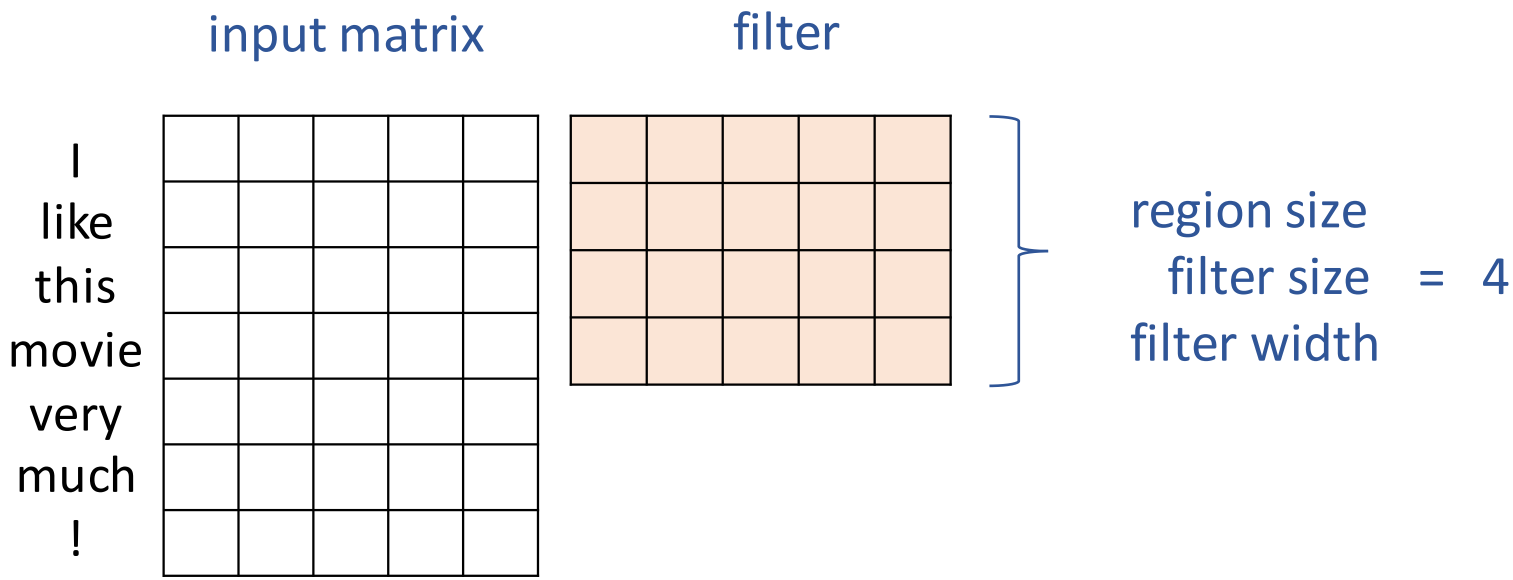
Convolutional neural networks

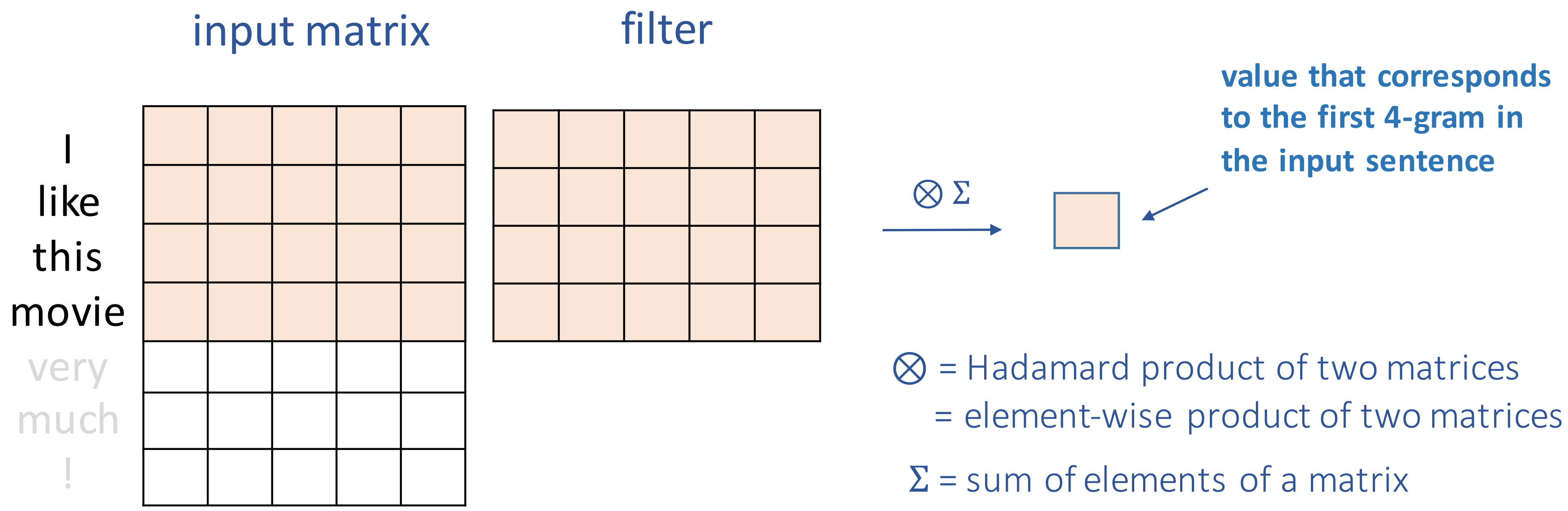
for sentence modeling

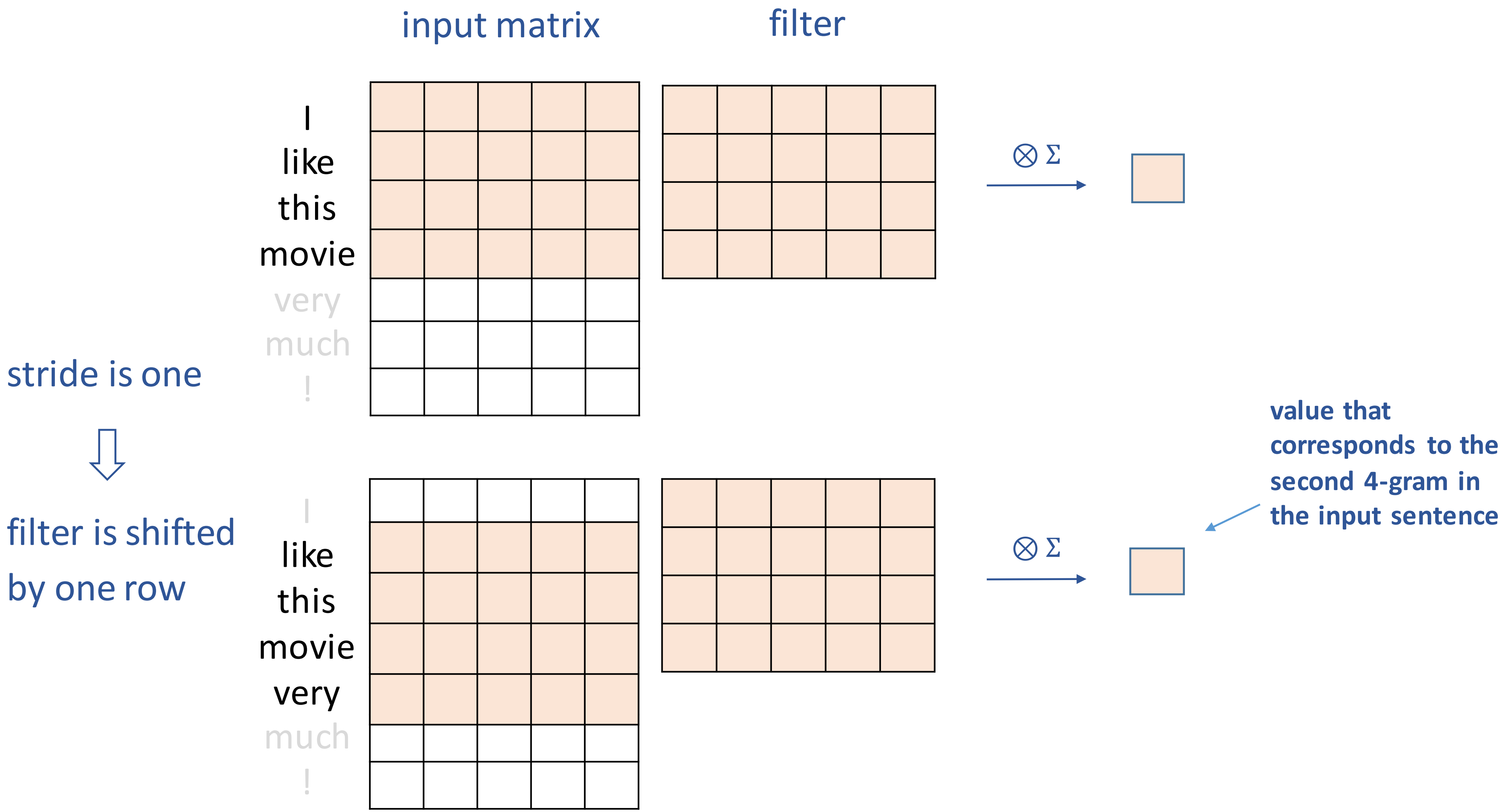
N. Kalchbrenner et al. “A Convolutional Neural Network for Modelling Sentences.” *ACL* (2014).

Y. Kim “Convolutional Neural Networks for Sentence Classification.” *EMNLP* (2014).

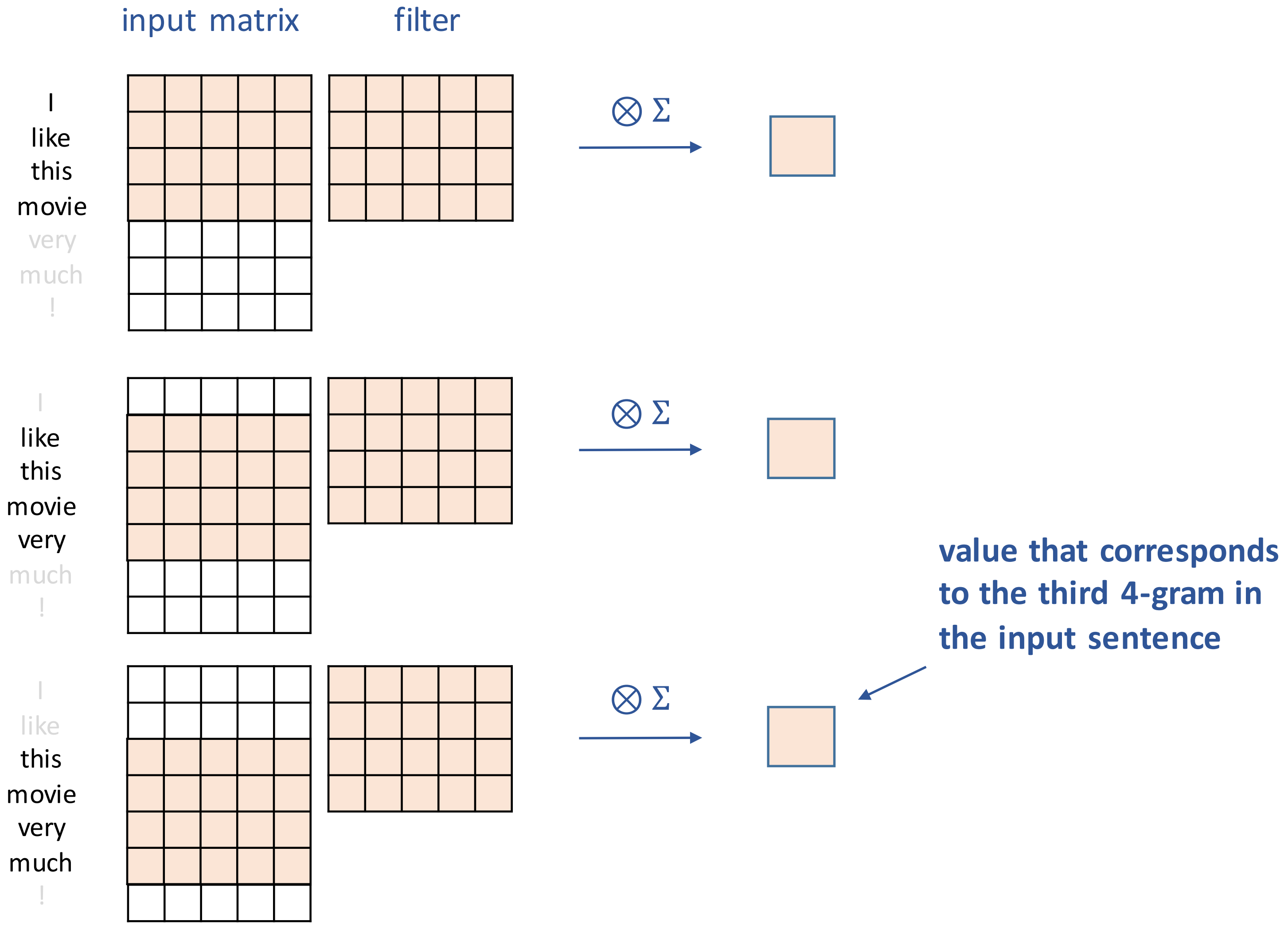
One-layer convolutional neural network

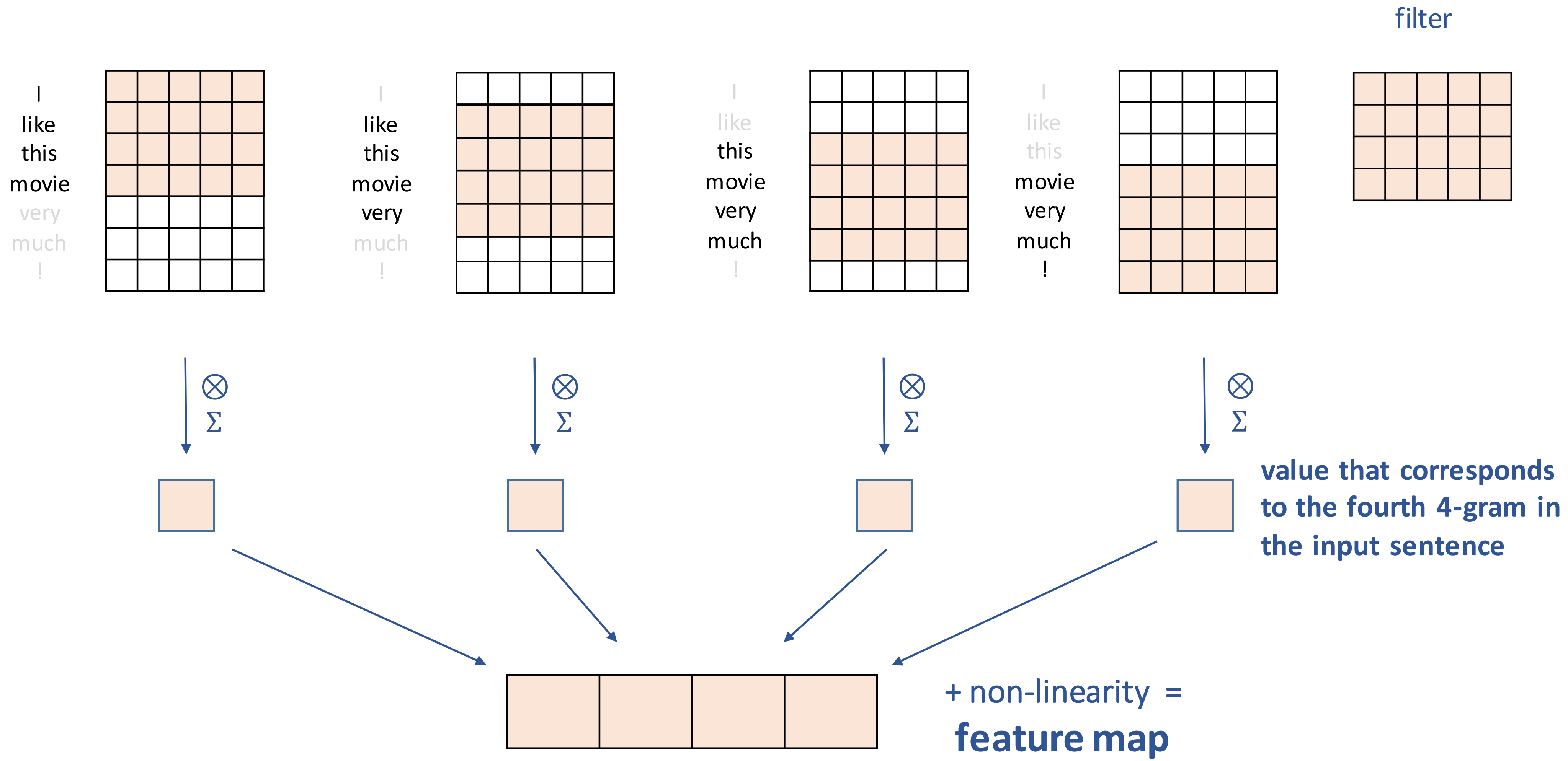


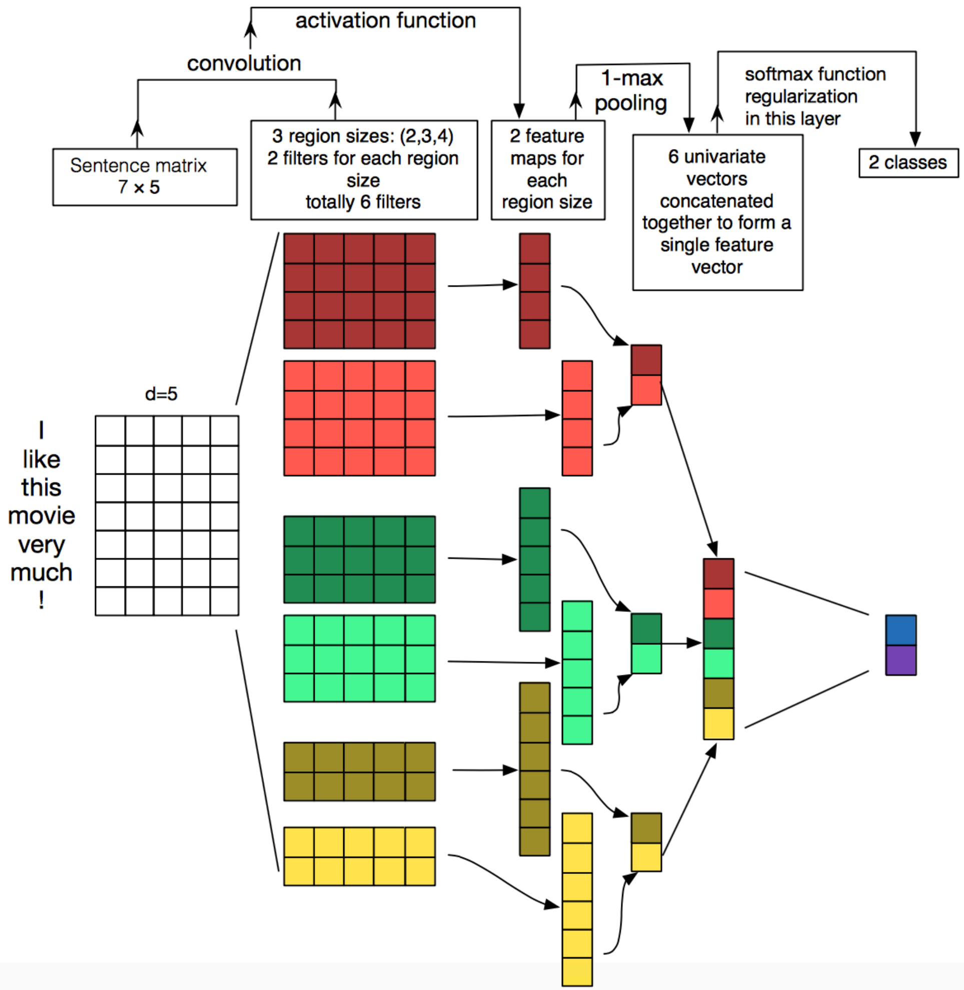

input matrix
filter
I
like
this
movie
very
much
!
⊗ Σ
value that corresponds to the first 4-gram in the input sentence
⊗ = Hadamard product of two matrices
= element-wise product of two matrices
Σ = sum of elements of a matrix



input matrix
filter
I
like
this
movie
very
much
!
⊗ Σ
stride is one
filter is shifted
by one row
value that
corresponds to the
second 4-gram in
the input sentence


input matrix
filter
I
like
this
movie
very
much
!
⊗ Σ
value that corresponds to the third 4-gram in the input sentence


filter
I
like
this
movie
very
much
!
value that corresponds to the fourth 4-gram in the input sentence
+ non-linearity =
feature map



One-channel convolutional neural network used in Kim (2014).
Figure taken from Zhang et al. (2015).

Properties of one-layer CNN

- CNN handles input sequences of varying length
- **CNN does not depend on external language-specific features such as dependency or constituent parse trees**
- CNN is sensitive to the order of the words in the sentence
- **Filters serve as feature detectors**
- **Convolving the same filter with the n-gram at every position in the sentence allows the features to be extracted independently of their position in the sentence**

CNN for MSC

MSC as a sentence classification task with a fixed sense inventory

Experimental setup: data

Corpora

- $MPQA_E$ (R&R)
- $EPOS_E$ and $EPOS_G$

  from EuroParl & OpenSubtitles (EPOS) heuristically tagged via cross-lingual sense projection

  in case of rare extractions for German: additional data from MVs with shared senses was added
- $MASC_E$: manually annotated subset of the multi-genre corpus MASC
- $TEST_G$: manually annotated instances from $EPOS_G$

Experimental setup: continued

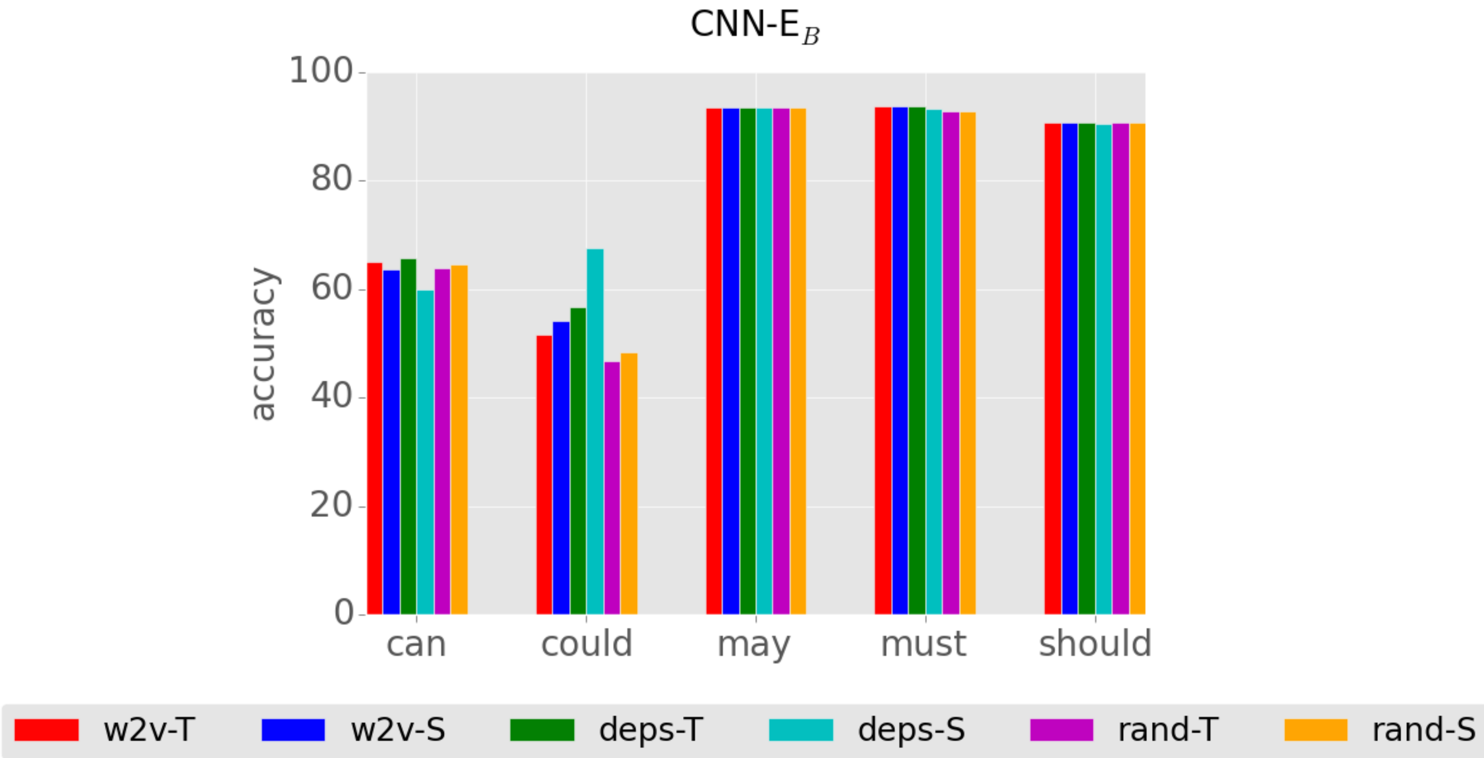
Hyperparameters (Zhang, 2015):

- **non-linearity:** ReLU
- **filter region sizes:** 3, 4, 5
- **number of filters per region size:** 100
- **dropout keep probability:** 0.5
- l_2 **regularization coefficient:** 10^{-3}
- **number of iterations:** 1001
- **mini-batch size:** 50
- **optimizer:** Adam optimization algorithm with learning rate 10^{-4}

Input representation: tuned and static versions of the following word vectors

- randomly initialized
- word2vec (Mikolov et al.)
- dependency-based (Levy et al.)

Impact of word vectors (E)



- train dataset: **balanced** 80% MPQA (R&R) + $EPOS_E$
- test dataset: (unbalanced) 20% MPQA
- accuracy with 5-fold CV

Comparison of CNN and baselines (E)

	can (3)	could (3)	may (2)	must (2)	should (2)	micro
BL_{random}	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	75.42	88.68	69.74
CNN	**65.78**	**67.50**	**93.57**	**93.82**	**90.77**	**79.29**

Classifiers trained on the **balanced** dataset. For every modal verb the best word vectors for it are used.

- train dataset: **balanced** 80% MPQA (R&R) + $EPOS_E$
- test dataset: (unbalanced) 20% MPQA
- accuracy with 5-fold CV

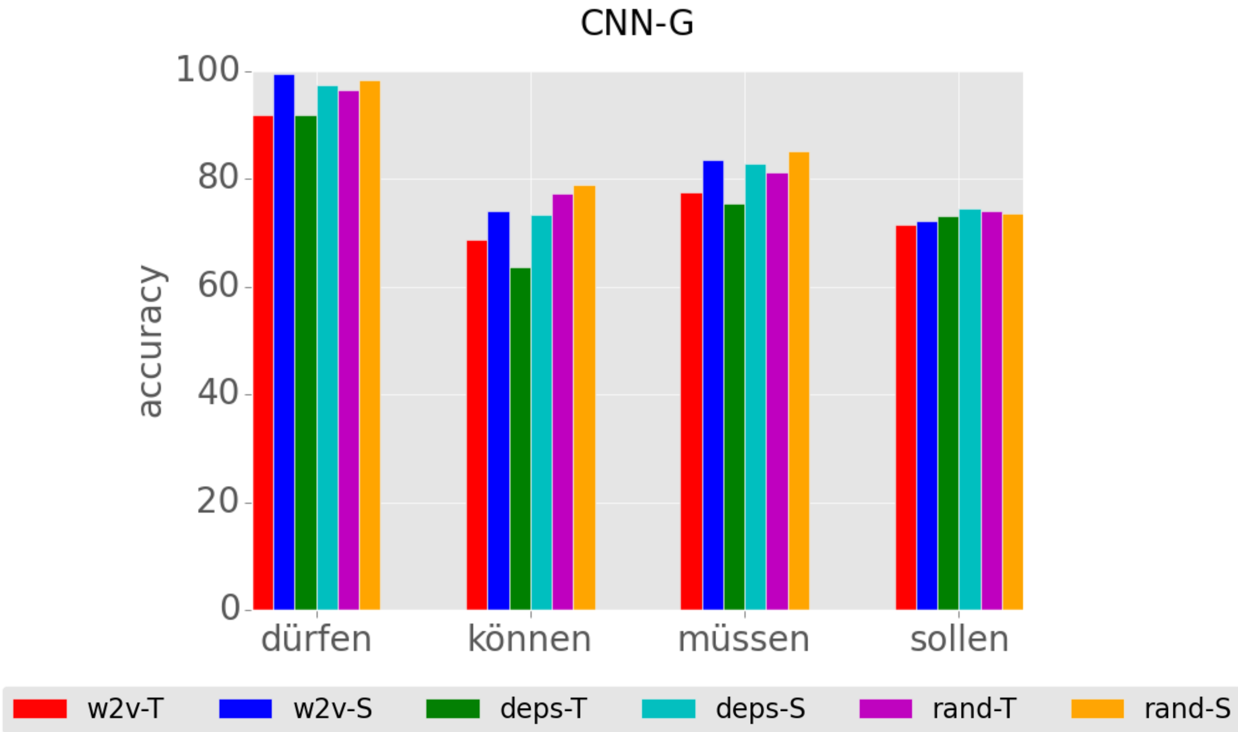
Comparison of CNN and baselines (E)

	can (3)	**could (3)**	**may (2)**	**must (2)**	**should (2)**	**micro**
$BL_{majority}$	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	**70.87**	**66.55**	93.49	**94.97**	90.59	**80.74**

Classifiers trained on the **unbalanced** dataset. For every modal verb the best word vectors for it are used.

- train dataset: **unbalanced** 80% MPQA (R&R) + $EPOS_E$
- test dataset: (unbalanced) 20% MPQA
- accuracy with 5-fold CV

Impact of word vectors (G)



- train dataset: **balanced** $EPOS_G$
- test dataset: $TEST_G$
- accuracy on the test dataset
- 1772 words from 10166 in vocabulary don't have pre-trained word2vec
- 2087 words from 10166 in vocabulary don't have pre-trained dep.-based vector

Comparison of CNN and baselines (G)

	dürfen	können	müssen	sollen	micro
BL_{random}	50.00	33.33	50.00	50.00	39.10
NN	80.30	48.89	74.63	49.75	60.00
CNN	**99.49**	**81.78**	**88.06**	**76.62**	**86.02**

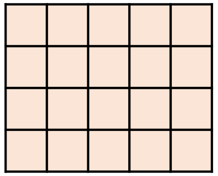
- train dataset: **balanced** $EPOS_G$
- test dataset: $TEST_G$
- accuracy on the test dataset

Visualizing what filters have learned

first sentence
second sentence
.
.
.
n-th sentence

\+

filter



⇒

max value of the first sentence
max value of the second sentence
.
.
.
max value of the *n-th* sentence

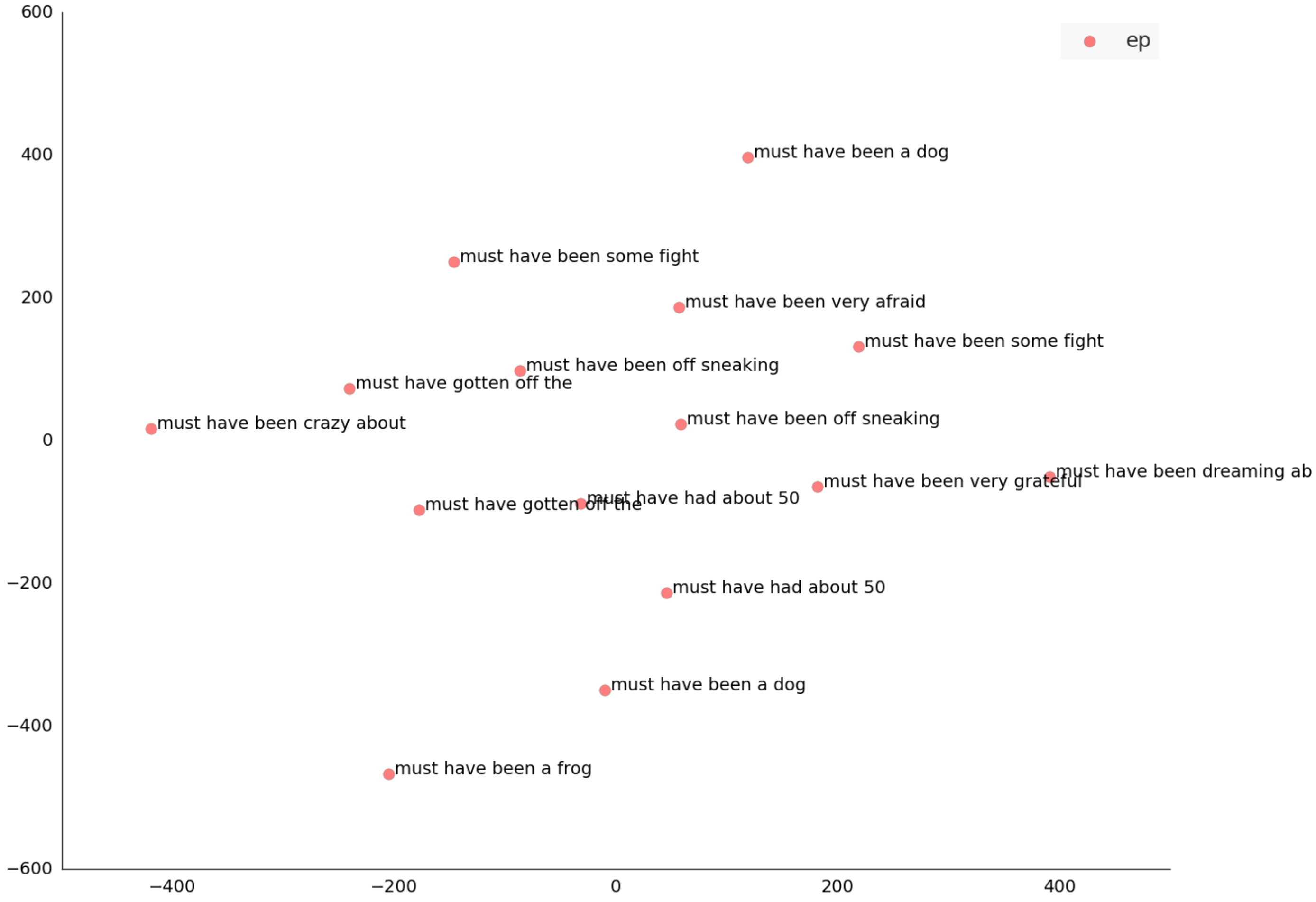
⇓

top 15 sentences w.r.t. the max value

⇓

n-gram from each sentence corresponding to the max value

Top 15 5-grams with respect to one filter illustrated in the embedded space



Feature detectors for *must*

feature	sense	example
past reading of the emb. verb	ep	**you must have been** out last night
non-past reading of the emb. verb	de	**we must take** further efforts
stative reading of the emb. verb	ep	**you must think** me a perfect tool
eventive reading of the emb. verb	de	**we must develop** a policy
passive construction	de	actual steps **must be taken**
negation	de	**we must not** fear
domain specific vocabulary	de	European parliament, present regulation, fisheries policy
telic clauses	de	to address these problems, to prevent both forum, to exert maximum influence
discourse markers	de	but, and (then)

Feature detectors for müssen and können

feature	sense	example
features that relate to observations on English		
attitude predicates	ep	believe, not know, tell me, have an idea, be afraid
adverbials	ep	possibly
conditionals	ep	if
counterfactual and negative polarity context	ep	bot be the case, how, ever
placeholders for propositions	ep	it
abstract concepts	ep	Idea, music, grades, application
indefinite subjects	ep	one
3rd person pronouns	ep	-
verb-object combinations for action that can be granted	de	use telephone
achievements (können only)	dy	present report

Other observations

Statistics

1) average distance of top ngrams from the modal verb
2) average distance of top ngrams which are on the left from the modal verb
3) as 2) but for ngrams on the right and ngrams starting with the modal

Observations

- there are no greater overall distances for German compared to English
- for German considerably more ngrams that include the modal verb, especially for epistemic readings of können, müssen, dürfen, but not for sollen
- strikingly larger distances to the left of the modal verb for epistemic readings compared to non-epistemic

Recap

- novel approach for multilingual MSC using a one-layer CNN
- CNN approach outperforms feature-based baselines
- CNN is able to learn meaningful structure from data
- CNN learns both known and previously unattested linguistic features for MSC and domain-specific concepts
- CNN learns linguistic and semantic features from flexible window regions without syntactic pre-processing
- CNN is easily adaptable to novel languages
- CNN allows for insightful model inspection, but this requires manual work

Word sense disambiguation

- if features CNN picks relate to semantic factors

  → CNN should be a good candidate for WSD

- features CNN picks relate to n-grams independent of their position in the sentence

  → CNN is flexible

  → can wider context be useful for WSD?

Comparison with the results from Rothe and Schütze (2015)

SensEval-3			
surrounding word	65.30		
local collocation	64.70	IMS (state-of-the art)	72.30
S_{naive} - product	62.20	IMS + S_{naive} - product	69.40
S - cosine	60.50	IMS + S - cosine	72.40
S - product	64.30	IMS + S - product	**73.60**
S - raw	63.10	IMS + S - raw	66.80
CNN	**67.90**	IMS + CNN	72.00

$w \ldots$ ambiguous word with k senses

$c \ldots$ centroid = sum of all w2v vectors of words in the sentence

$s^{(j)} \ldots$ embedding of the j-th synset of w

$$\text{S-cosine} = \langle \cos(c, s^{(1)}), \ldots, \cos(c, s^{(k)}) \rangle$$

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

$$\text{S-raw} = \langle c_1, \ldots, c_n, \ldots, s_1^{(k)}, \ldots, s_n^{(k)} \rangle$$

AutoExtend

- for the sentence representation all constituent words are available
- rich knowledge about the target word

CNN

- for sentence representation all constituent word are available
- without any knowledge of the target word
- flexibility (wider context) in picking relevant n-grams

Future work

Feature work on WSD

- tune hyperparameters
- use more data
- use deeper network

Feature work in general

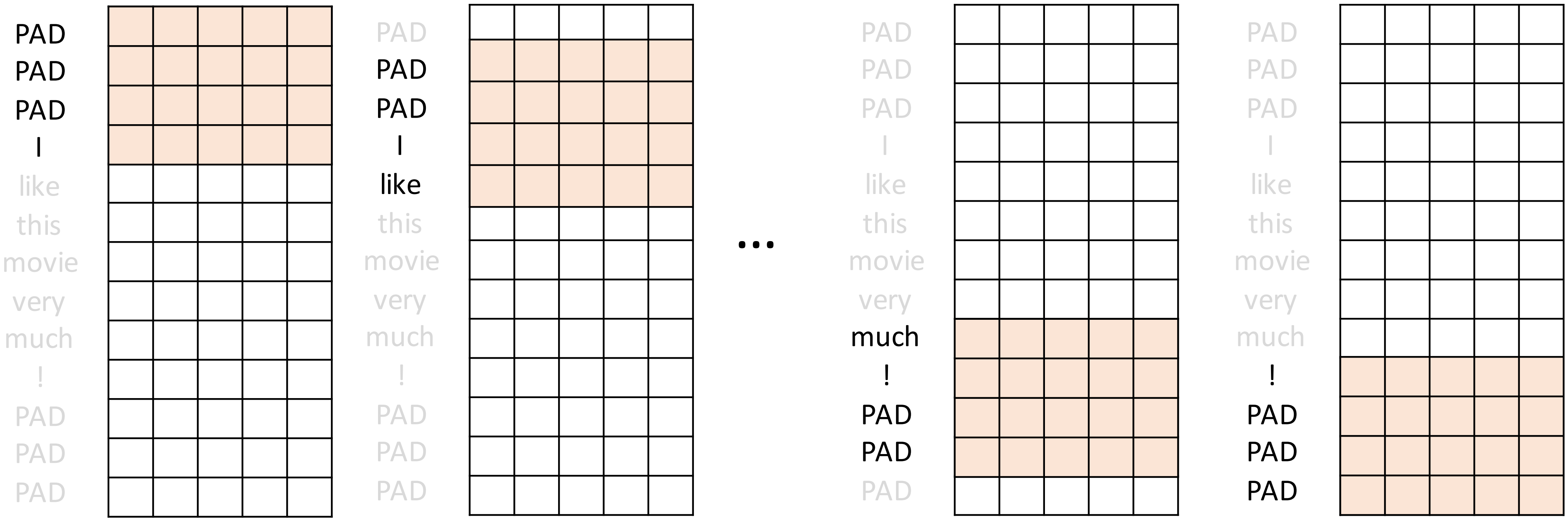
- extraction of opinion entities: opinion expressions, their holders and targets
- implicit sentiment: where MSC plays role
  - **planned (positively): should, must + deontic**
  - **apprehended (negative): should not + deontic**
  - **disliked or forbidden (negative): may not +deontic**
  - **desired (positive): should + deontic**

Thank you for your attention!

References

- N. Kalchbrenner et al. "A Convolutional Neural Network for Modelling Sentences." *ACL* (2014).
- Y. Kim "Convolutional Neural Networks for Sentence Classification." *EMNLP* (2014).
- O. Levy and Y. Goldberg. "Dependency-Based Word Embeddings." *ACL* (2014).
- T. Mikolov et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems 26 (2013).*
- S. Rothe and H. Schütze. "AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes." *ACL* (2015).
- J. Ruppenhofer and I. Rehbein. Yes we can!? Annotating the senses of English modal verbs. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)* (pp. 24-26). (2012)
- Y. Zhang and W. Byron. "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification." *CoRR* abs/1510.03820 (2015): n. pag.
- M. Zhou, A. Frank, A. Friedrich and A. Palmer. Semantically Enriched Models for Modal Sense Classification. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)* (p. 44) (2015) .

Narrow and wide convolution



s = sentence length

m = filter region size

narrow convolution ⇒ feature map size equals to **s-m+1**

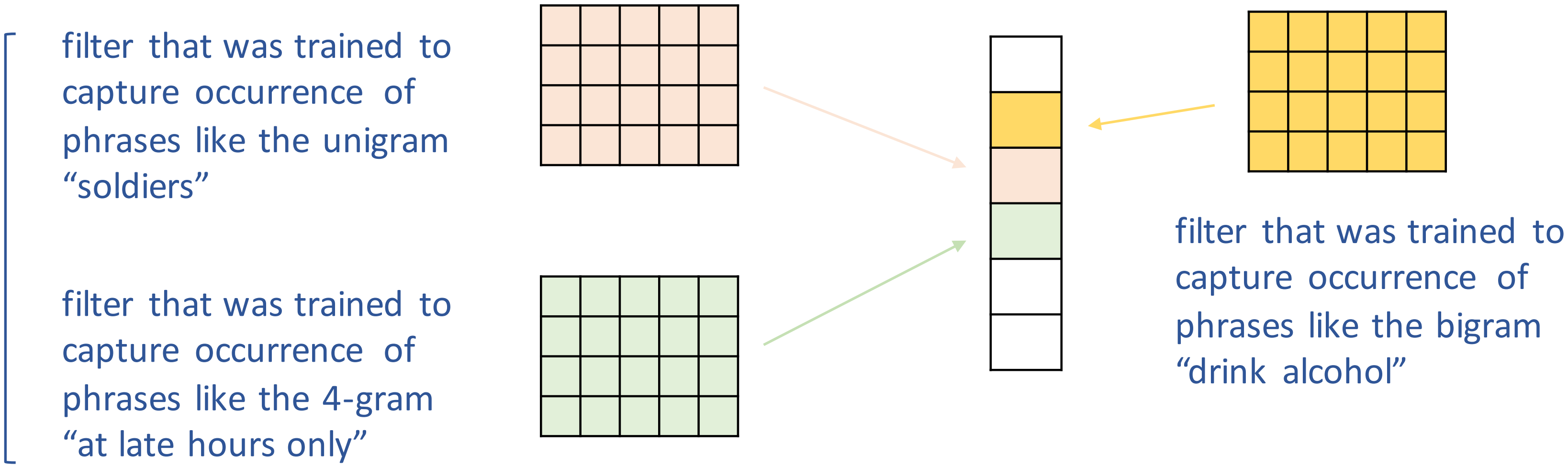
wide convolution ⇒ feature map size equals to **s+m-1**

Will a one-layer CNN be sufficient?

Soldiers can drink alcohol until they fall over. (dynamic-capability)

Soldiers can drink alcohol at late hours only. (deontic-permission)

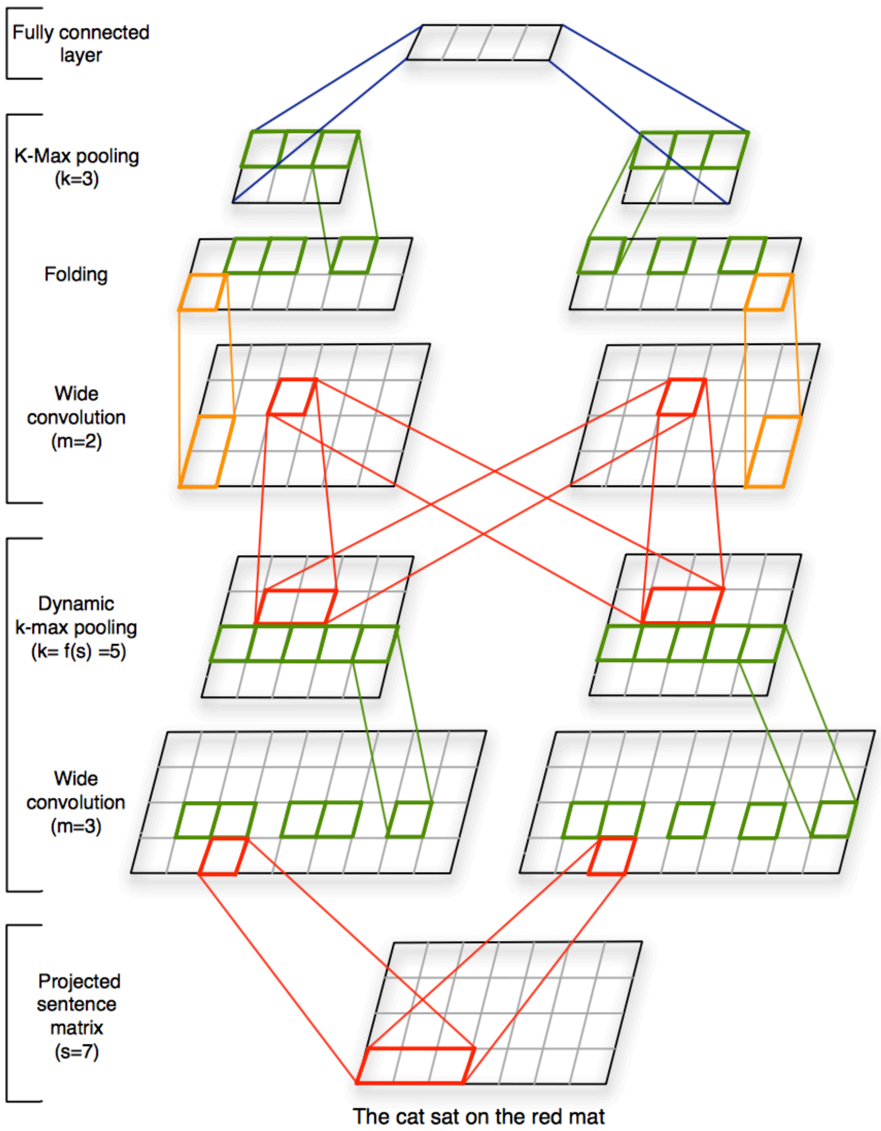
how can does this miracle come about?



Dynamic Convolutional Neural Network (DCNN)

Kalchbrenner et al. (2014)

- Wide type of convolution
- One-dimensional filter to each row of the input matrix
- **Stacked convolutional layers**
- ***k*-max pooling**
- *k* is a function of the length of the sentence and the depth of the network



Train-test configurations

	train	test
English	80% $MPQA_E$ (R&R) + $EPOS_E$ +/- balancing	a. 20% $MPQA_E$ (R&R) w/ 5-fold CV b. $MASC_E$
German	$EPOS_G$	$TEST_G$

Impact of word vectors (E)

	can (3)	could (3)	may (2)	must (2)	should (2)
w2v-static	65.02	51.67	**93.57**	**93.82**	**90.77**
w2v-tuned	63.73	54.17	**93.57**	**93.82**	**90.77**
deps-static	**65.78**	56.67	**93.57**	**93.82**	**90.77**
deps-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**

- train dataset: **balanced** 80% MPQA (R&R) + EPOS_E
- test dataset: (unbalanced) 20% MPQA
- accuracy with 5-fold CV

MASC

	test dataset	training dataset: balanced/unbalanced	
MaxEnt	$MPQA_E$	BA (74.88)	UBA (78.01)
	$MASC_E$	BA (3/19)	UBA (15/19)
CNN	$MPQA_E$	BA (79.92)	UBA (80.74)
	$MASC_E$	BA (13/19)	UBA (3/19)

Balanced vs. unbalanced training when evaluated on $MPQA_E$ and $MASC_E$ for CNN and MaxEnt.

balanced (BA) training data	CNN (19/19)	MaxEnt (0/19)
unbalanced (UBA) training data	CNN (12/19)	MaxEnt (7/19)

Difference between CNN and MaxEnt trained on $MPQA_E$ +EPOS and evaluated on MASC.

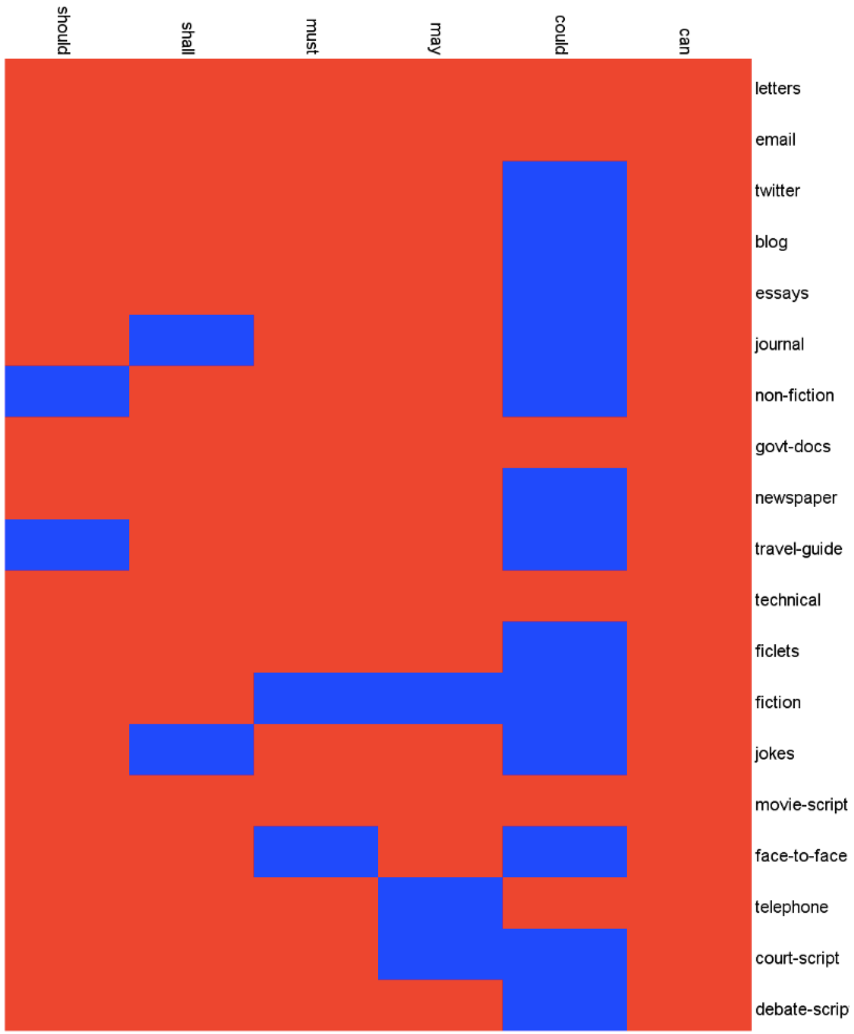
Appendix: MASC

CNN_BA - CNN_UBA	must	may	can	could	should	
letters	0	0	0	-7,692307692	0	-0,970873786
email	0	0	4,444444444	-4,545454545	-5	0
twitter	0	0	1,666666667	0	0	0,877192982
blog	-20	0	0	7,692307692	0	0
essays	0	0	0	20	0	4,301075269
journal	0	-6,25	0	11,11111111	0	1,075268817
non-fiction	0	0	0	16,66666667	0	0,884955752
govt-docs	0	0	0	12,82051282	0	4,464285714
newspaper	0	0	0	20	0	7,058823529
travel-guides	0	0	-1,886792453	14,28571429	0	0
technical	0	0	0	-4,166666667	0	-1,25
ficlets	-11,11111111	0	-2,325581395	29,72972973	0	8,411214953
fiction	-6,666666667	0	-6,451612903	28,88888889	0	7,407407407
jokes	11,11111111	0	-5,454545455	21,05263158	0	2,127659574
movie-script	0	20	0	6,25	0	2,083333333
face-to-face	0	0	1,470588235	28,94736842	0	8,163265306
telephone	0	0	0	18,18181818	0	4,545454545
court-transcript	0	3,846153846	-4,411764706	6,25	0	4,255319149
debate-transcript	-16,66666667	0	1,515151515	25	0	-0,709219858
micro-avg	-2,083333333	0,380228137	-0,574712644	15,32846715	-0,398406374	**2,840610205**
macro-avg	-2,407407407	0,92611336	-0,601760318	13,18275371	-0,263157895	

Appendix: MASC

CL2BA - CL2UBA	must	may	can	could	should	
letters	-12,5	0	-8,7	-7,69	0	-7,77
email	0	0,01	-6,67	4,55	-10	-3,7
twitter	-10	0	0	0	-5	-1,75
blog	0	0	-17,78	30,77	0	-3,77
essays	-11,11	0	-2,94	15	-9,09	0
journal	0	0	-13,16	16,66	0	-2,15
non-fiction	0	0	-5,88	16,66	0	-2,66
govt-docs	-10	0	-5	15,39	-18,75	0,89
newspaper	-16,66	0	-13,05	6,66	-9,08	-3,53
travel-guides	0	0	-3,77	14,28	0	-1,12
technical	0	0	0	-12	0	-3,71
ficlets	-11,11	0	-4,65	0	0	-2,81
fiction	0	7,15	-9,38	20	0	5,15
jokes	11,11	0	-10,91	15,79	-22,22	-4,26
movie-script	0	0	-7,84	-18,75	7,69	-6,18
face-to-face	50	0	-7,35	2,63	-2,86	-2,72
telephone	0	0	-7,69	-9,09	-21,05	-13,63
court-transcript	0	0	-1,47	15,62	0	2,9
debate-transcript	0	0	-1,51	0	0	-1,13
micro-avg,	-2,78	0,38	-6,43	7,52	-5,58	**-2,16**
macro-avg,	-0,54	0,37	-6,73	6,66	-4,76	

Appendix: MASC



Appendix: MASC

CNN_BA - CL2BA	must	may	can	could	should	
letters	12,5	-0,001818182	14,49434783	-15,38384615	0	8,736893204
email	8,333333333	-0,002222222	17,77777778	-9,093636364	5	7,407037037
twitter	-10	0	5	-15,38384615	5	0,872982456
blog	-20	5	22,22666667	-15,38384615	-0,002352941	7,77
essays	22,22	-0,003157895	8,819411765	5	9,089090909	7,530430108
journal	-0,003333333	-12,5	13,15578947	5,56	0	4,298817204
non-fiction	0	0,001724138	7,354117647	-16,66333333	0	3,541415929
govt-docs	20	-3,69962963	5	-5,131025641	18,75	2,681428571
newspaper	33,33	0	13,04521739	13,33666667	9,088181818	11,76705882
travel-guides	0	0	3,768867925	14,28714286	0	3,369662921
technical	0	0	-0,001428571	10,83333333	-0,001111111	3,49
ficlets	-0,002222222	0	13,95209302	10,80567568	-0,004615385	9,347570094
fiction	-13,33666667	-7,147142857	31,0483871	4,443333333	-3,333333333	5,729259259
jokes	-0,001111111	0	10,91181818	-0,002105263	22,22	8,513404255
movie-script	0,003333333	0	17,25	-12,5	-7,691538462	5,895
face-to-face	-50	0	20,58588235	13,16263158	2,86	12,92292517
telephone	0	0	15,38538462	9,093636364	21,05	15,90545455
court-transcript	33,33	0,004615385	17,64941176	-9,375	-0,001111111	17,81106383
debate-transcript	-33,33333333	0	16,6630303	12,5	0	2,453546099
micro-avg	1,392222222	-1,137984791	13,71977011	1,616885645	4,784621514	**7,076670174**
macro-avg	4,665555556	-0,960927961	13,37614607	0,00030425	4,319116336	

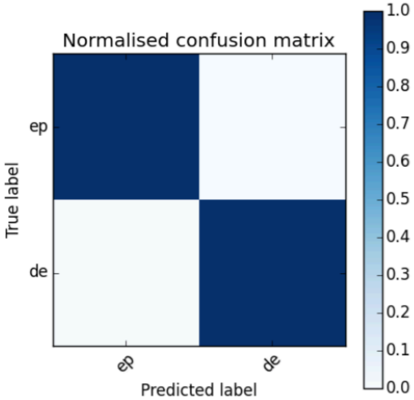
Appendix: MASC

CNN_UBA - CL2UBA	must	may	can	could	should	
letters	0	-0,001818182	5,794347826	-15,38153846	0	1,93776699
email	8,333333333	0,007777778	6,663333333	0,001818182	0	3,707037037
twitter	-20	0	3,333333333	-15,38384615	0	-1,754210526
blog	0	5	4,446666667	7,693846154	-0,002352941	4
essays	11,11	-0,003157895	5,879411765	0	-0,000909091	3,229354839
journal	-0,003333333	-6,25	-0,004210526	11,10888889	0	1,073548387
non-fiction	0	0,001724138	1,474117647	-16,67	0	-0,003539823
govt-docs	10	-3,69962963	0	-2,561538461	0	-0,892857143
newspaper	16,67	0	-0,004782609	-0,003333333	0,008181818	1,178235294
travel-guides	0	0	1,885660377	14,28142857	0	2,249662921
technical	0	0	-0,001428571	3	-0,001111111	1,03
ficlets	-0,001111111	0	11,62767442	-18,92405405	-0,004615385	-1,87364486
fiction	-6,67	0,002857143	28,12	-4,445555556	-3,333333333	3,471851852
jokes	-0,002222222	0	5,456363636	-5,264736842	0	2,125744681
movie-script	0,003333333	-20	9,41	-37,5	-0,001538461	-2,368333333
face-to-face	0	0	11,76529412	-13,15473684	0	2,039659864
telephone	0	0	7,695384615	-18,17818182	0	-2,27
court-transcript	33,33	-3,841538461	20,59117647	-0,005	-0,001111111	16,45574468
debate-transcript	-16,66666667	0	13,63787879	-12,5	0	2,032765957
micro-avg,	0,695555556	-1,138212928	7,864482759	-6,191581509	-0,396972112	**2,076059968**
macro-avg,	6,532962963	-1,517041322	7,247906384	-6,522449459	-0,177725769	

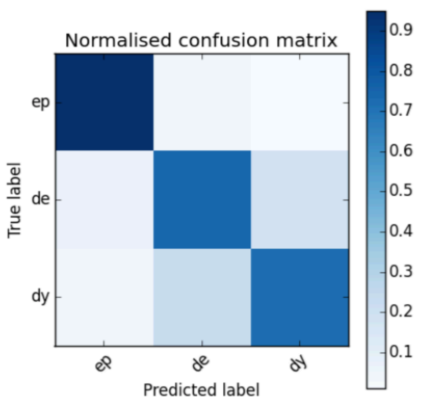
Comparison of CNN and baselines (G)

	dürfen	**können**	**müssen**	**sollen**	**micro**
BL_{random}	50.00	33.33	50.00	50.00	39.10
NN	80.30	48.89	74.63	49.75	60.00
CNN	**99.49**	**81.78**	**88.06**	**76.62**	**86.02**

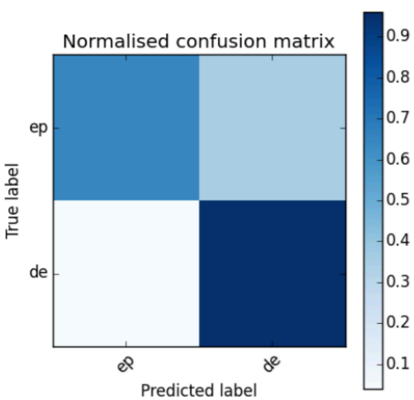
- train dataset: **balanced** $EPOS_G$
- test dataset: $TEST_G$
- accuracy on the test dataset



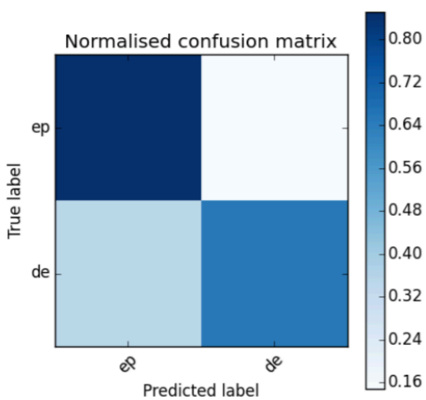
(a) dürfen



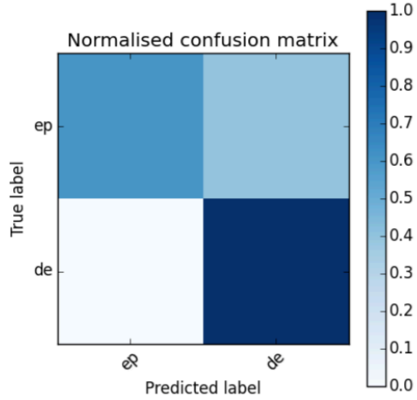
(b) können



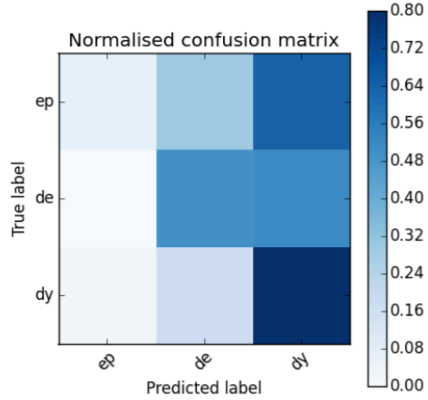
(c) müssen



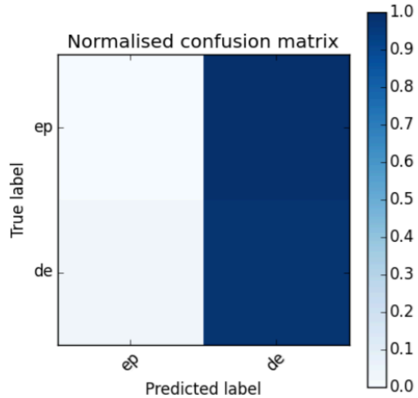
(d) sollen



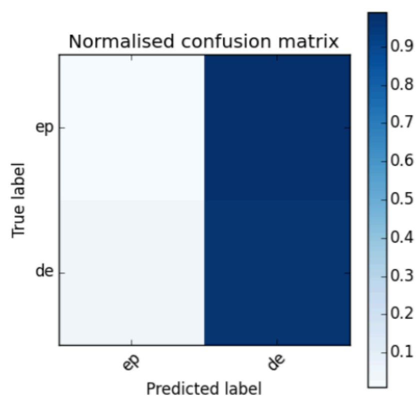
(a) dürfen



(b) können



(c) müssen



(d) sollen

CNN (German)

NN (German)

Appendix: number of instances

	CL^{-b}_{ME} train			CL^{+b}_{ME} train			MPQA test		
	ep	de	dy	ep	de	dy	ep	de	dy
must	806	949	0	870	870	0	5	34	0
may	1055	956	0	999	1000	0	25	3	0
can	151	248	362	250	250	250	1	17	60
could	160	55	97	94	94	94	36	2	10
should	171	355	0	250	250	0	5	52	0
shall	0	14	6	0	15	15	0	2	1

Appendix: number of instances (G)

	ep	de	dy
dürfen	1000	1000	0
können	1000	1000	1000
müssen	1000	1000	0
sollen	1000	1000	0

ep	de	dy
98	100	0
100	47	100
34	100	0
101	100	0

- train dataset: **balanced** EPOS_G
- test dataset: TEST_G