Revisiting neural relation classification in clinical notes with external information

Simon Šuster, Madhumita Sushil and **Walter Daelemans**
Computational Linguistics & Psycholinguistics Research Center,
University of Antwerp, Belgium
firstname.lastname@uantwerpen.be

Abstract

Recently, segment convolutional neural networks have been proposed for end-to-end relation extraction in the clinical domain, achieving results comparable to or outperforming the approaches with heavy manual feature engineering. In this paper, we analyze the errors made by the neural classifier based on confusion matrices, and then investigate three simple extensions to overcome its limitations. We find that including ontological association between drugs and problems, and data-induced association between medical concepts does not reliably improve the performance, but that large gains are obtained by the incorporation of semantic classes to capture relation triggers.

1 Introduction

The extraction of relations from clinical notes is a fundamental clinical NLP task, crucial to support automated health care systems and to enable secondary use of clinical notes for research (Wang et al., 2017). In clinical relation extraction, the 2010 i2b2/VA challenge dataset has been by far the most widely used. Three categories of relations are annotated in discharge summaries: those between medical treatments and problems (**TrP**)[1], between tests and problems (**TeP**)[2] and between pairs of problems (**PP**)[3] (Uzuner et al., 2011). Many systems participating in the shared task used carefully crafted syntactic and semantic features, sometimes in combination with rules (Grouin et al., 2010; Rink et al., 2011). Recently, neural network approaches have been applied to this task, where they serve as feature extractors, with a softmax layer for classification. In this case, human-engineered or external features are usually not included. Two examples on which we base our work are Sahu et al. (2016) and Luo et al. (2017), who achieve results similar to or better than the best-scoring approaches participating in the i2b2 challenge. They use convolutional neural networks, in which a convolutional unit processes a piece of text segment (SegCNN) in a sliding window manner, and then applies a max-pooling operation to provide the hidden features. In Sahu et al. (2016), the unit of text is simply a sentence, and the CNN constructs a global representation. On the other hand, Luo et al. (2017) argue that since multiple relations can occur in a single sentence, one representation is not sufficient. Therefore, they break the sentence into segments, so the encoding and the pooling operations apply to one segment at a time. Each sentence consists of five segments: tokens preceding the first concept c_1; c_1 itself; tokens between c_1 and c_2; concept c_2; and the tokens following it. This idea is related to *dynamic pooling*, known from previous event extraction work on the ACE 2005 dataset (Chen et al., 2015). More generally, the extension of neural networks with background information have been studied, inter alia, for text categorization, natural language inference, and entity and event extraction (K. M. et al., 2018; Yang and Mitchell, 2017).

In our work, we aim to boost the performance of a SegCNN classifier by first identifying its weakest points in a confusion matrix analysis, and then addressing these with external linguistic and domain features. We observe as much as a 6 point improvement in % F1 by a simple addition of semantic classes; a modest improvement with PMI features for PP relations; and no effect when adding association information between drugs and problems. We make the code, which is a modification of Luo et al. (2017)'s implementation of segment convolutional neural networks, available at https://github.com/SimonSuster/seg_cnn.

[1] Tr[A|C|I|NA|W]P: treatment {administered for, causes, improves, not administered because of, worsens} a problem.

[2] Te[C|R]P: test {conducted for, revealed} a problem.

[3] PIP: problem indicates a medical problem.

g\s	None	TrAP	TrCP	TrIP	TrNAP	TrWP
None	**980**	86	15	3	7	0
TrAP	139	**423**	5	3	3	0
TrCP	48	27	**69**	0	0	0
TrIP	11	12	1	**16**	0	0
TrNAP	11	24	3	0	**7**	0
TrWP	11	16	5	4	1	**4**

(a) TrP relations.

g\s	None	TeCP	TeRP
None	**575**	17	294
TeCP	41	**52**	36
TeRP	89	9	**612**

(b) TeP relations.

g\s	None	PIP
None	**2544**	135
PIP	122	**343**

(c) PP relations.

Table 1: Confusion matrices for different relation categories of the base SegCNN. The first diagonal represents the number of correctly classified relations, and is shown in bold. The colored cells highlight low sensitivity (blue), hallucinating relations (green) and confusable relations (orange).

2 Analysis of limitations

To better understand the limitations of a SegCNN extractor, we analyze its results with confusion matrices. In Table 1, we use color coding to point to three types of challenges: a) **poor sensitivity** (blue cells), which are errors due to the classifier's conservativeness in proclaiming a relation; b) **"hallucinating" relations** (green), which are precision errors where relations should not be identified; and c) **confusable relations** (orange), where we see that the TrCP relation is often classified as TrAP (27/69 times), and similarly for the other treatment-problem relations. This is especially true for the less frequent relations TrNAP and TrWP, where the correct predictions are outnumbered by the cases wrongly predicted as TrAP. The TrAP predictions by the system account for the most mistakes. We can see from the number of a) and b) errors on the TrP relations—76% of all mistakes made by the model—that identifying the presence of a relation is more challenging than type classification of relations, cf. Rink et al. (2011). Similar observations can be made about the test-problem relations. For example, TeCP is frequently confused with TeRP (36), and the TeRP type is often hallucinated (294). Overall, determining the presence of a relation is more difficult than discriminating between TeCP and TeRP as 91% of mistakes are only due to detection. This number is higher here than for TrP relations since we are dealing with a smaller number of relation types, which causes less confusion in class assignment. For problem-problem relations, the matrix shows the model is somewhat more likely to predict the relation spuriously than to miss the relation.

In a qualitative analysis, we find that relations are often unrecognized in sentences with several (coordinated) concepts:

(1) *she also had climbing bilirubin [. . .] and was started on zosyn*$_{tr}$ *for suspected biliary obstruction and ascending cholangitis*$_{pr}$ *coverage .* (gold: TrAP)

Relations can be hallucinated especially when two concepts may seem to be associated, but the knowledge of syntax or the domain tells us they are not:

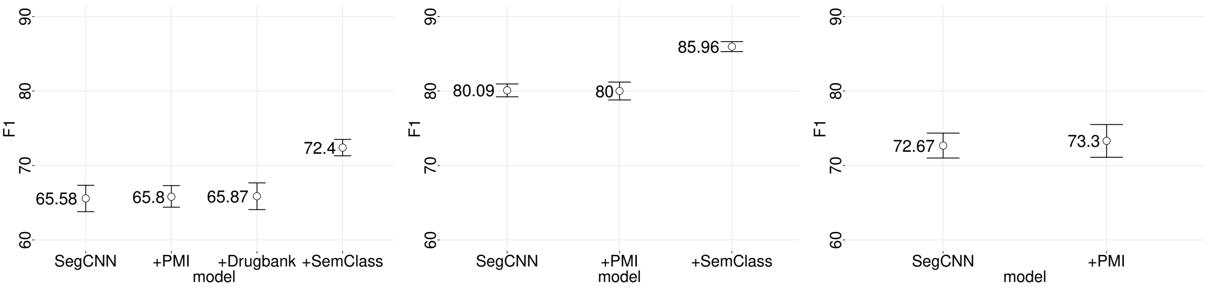
(2) *the patient was treated with tylenol orally*$_{tr}$ *as well as ativan for anxiety*$_{pr}$ *that she had about going home* (gold: none)

Here, medical knowledge of compatibility between drugs and problems could help, e.g. that tylenol is not indicated for anxiety, but ativan is. In the following example, the classifier wrongly predicts TeCP, although there is a clear cue for the correct relation TeRP in the predicate ("found"):

(3) *during initial evaluation*$_{te}$ *for a coronary artery bypass graft , 80% to 90% of the right coronary artery stenosis*$_{pr}$ *was found*

3 Addressing the limitations

To deal with poor sensitivity and hallucinated relations mentioned above, we introduce simple domain knowledge in the form of association between a pair of concepts. We collect the association information either from an ontology (§ 3.1) or induce it from the data (§ 3.2). To increase the discriminatory power of the extractor to differentiate between the relations, we incorporate a semantic class feature which could give the classifier an explicit cue about the presence of a relation (§ 3.3).



(a) TrP

(b) TeP

(c) PP

Figure 1: Results per relation category in percentage F1. The reported scores are averaged over 20 runs, and the 95% confidence intervals are shown.

3.1 Drug-problem association (Drugbank)

We use Drugbank (Wishart et al., 2017) to obtain a compatibility score between a drug treatment and a problem. We create a mapping from all drug names, synonyms and product names, to their indications. We also extract a mapping between drugs and their adverse reactions. In this way, we obtain 71,683 drug names, 3108 indications and 1163 adverse reactions. If there is a match for an observed treatment-problem pair in the drug-indication mapping, we simply assign a value of 1 (and scale it, as explained in Appendix) and -1 otherwise. Consider the example where we consider creating a relation between $\underline{\textit{neurontin}}_{tr}$ and $\underline{\textit{seizure history}}_{pr}$. In the indication for neurontin from Drugbank, seizures are mentioned as a possible medical problem, so this type of information could serve as background evidence for the classifier. The adverse drug effects represent a separate feature and are included in the same way. Due to low coverage of the drug-problem features for the treatment-problem concept pairs in the data (416 pairs are found, out of 7699), we also investigate a more general, data-induced approach, described next.

3.2 Concept-concept association (PMI)

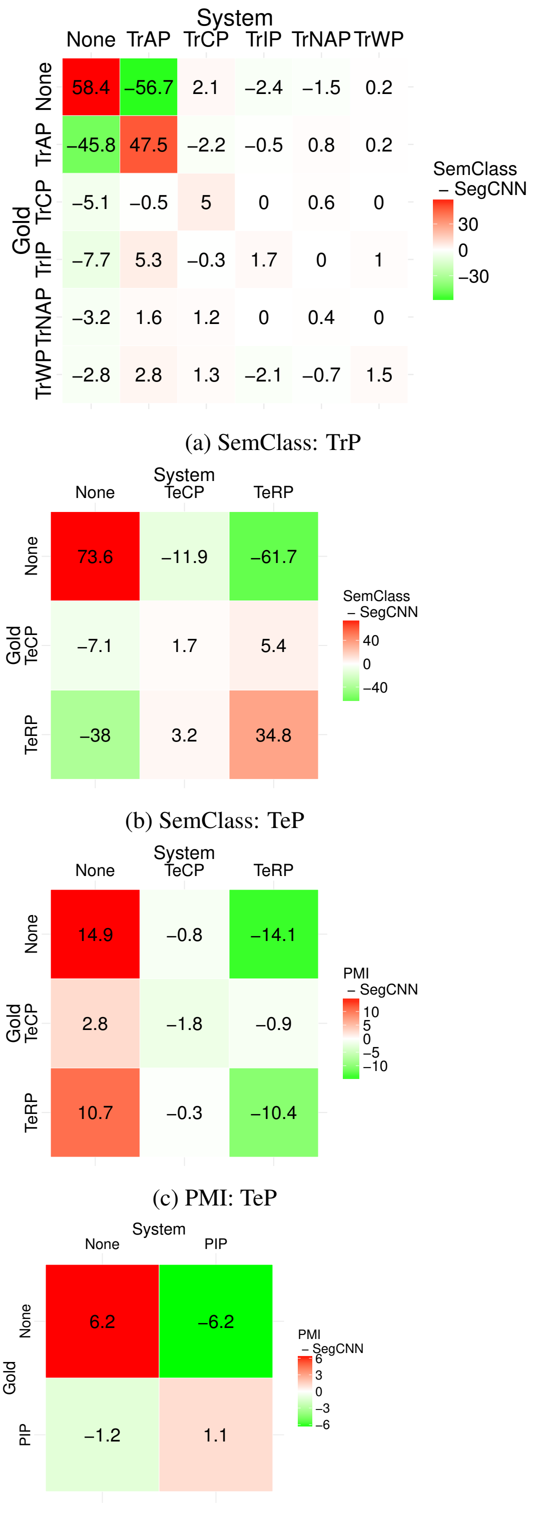
We obtain association scores for concept pairs in all relation types by estimating a pointwise mutual information (PMI) model on a large corpus. We use the MIMIC-III corpus (Johnson et al., 2016) to compute the PMI for the co-occurring concepts. We first recognize clinical concepts in MIMIC-III using CLAMP (Soysal et al., 2017), and use Ucto (Van Gompel et al., 2012) for preprocessing. We then collect the counts, where two concepts are taken as co-occurring if they are mentioned in the same sentence, irrespective of the ordering. If found, we remove any determiners and pronouns. The concept type identified by CLAMP is appended to its mention. For a concept pair in our data, we perform a type-sensitive and order-insensitive lookup. In case of no match, we back-off by gradually removing up to two leftmost tokens. We find that the coverage lies between 68–82% depending on the relation category and the dataset split, and that the highest coverage applies for PP relations. The concept-concept association for relation extraction has been studied previously by Demner-Fushman et al. (2010) and de Bruijn et al. (2011), who used Medline® as the resource, whereas we achieved better results and coverage on the development set with MIMIC-III than Medline®.

3.3 Semantic classes

The semantic classes can provide cues about the relation types present in the sentence and facilitate distinguishing between different TrP and TeP relations[4]. We obtain the classes with WordNet (Miller, 1995) and an online thesaurus[5]. This was a manual process, in which we looked up the synonyms for all relation type names. For the seven TrP and TeP relation types, a hundred lexical triggers were obtained in total. For example, {*show*, *reveal*, *display*…} belong to the "revealing" class indicative of the TeRP relation. Lexical triggers are matched to their semantic classes if they occur in the non-concept sentence segments. We find that for TrP relations, matching only with the middle segment works best, but for TeP, the preceding, middle and succeeding segments work best.

[4] We do not use semantic classes for PP since there is only one relation type, PIP.

[5] en.oxforddictionaries.com



Gold \ System	None	TrAP	TrCP	TrIP	TrNAP	TrWP
None	58.4	–56.7	2.1	–2.4	–1.5	0.2
TrAP	–45.8	47.5	–2.2	–0.5	0.8	0.2
TrCP	–5.1	–0.5	5	0	0.6	0
TrIP	–7.7	5.3	–0.3	1.7	0	1
TrNAP	–3.2	1.6	1.2	0	0.4	0
TrWP	–2.8	2.8	1.3	–2.1	–0.7	1.5

(a) SemClass: TrP

Gold \ System	None	TeCP	TeRP
None	73.6	–11.9	–61.7
TeCP	–7.1	1.7	5.4
TeRP	–38	3.2	34.8

(b) SemClass: TeP

Gold \ System	None	TeCP	TeRP
None	14.9	–0.8	–14.1
TeCP	2.8	–1.8	–0.9
TeRP	10.7	–0.3	–10.4

(c) PMI: TeP

Gold \ System	None	PIP
None	6.2	–6.2
PIP	–1.2	1.1

(d) PMI: PP

Figure 2: A comparison of counts between a SegCNN and a model using either semantic classes or PMI features, for different relation categories.

4 Results

In our experiments, we use different data splits from those used in Luo et al. (2017) to increase the size of the training part and to also create a development set. The details, including the experimental setting, can be found in the Appendix. For the results using the vanilla SegCNN, we retrain the original models by Luo et al. (2017) and report their performance on our data splits. This gives us the results which are a few points lower on TrP and TeP relations, but also few points higher on PP relations, than the results reported in their paper.

We show the results in Figure 1, where % F1 is reported for different relation categories. Overall, the highest scores are achieved on TeP relations. The addition of semantic classes helps the most, with an improvement of almost 7 points over SegCNN for TrP, and 6 points for TeP relations. We think the advantage comes from the fact that the relation triggers are represented explicitly as the input to the classifier, whereas in the case of the base SegCNN, the classifier can only rely on a dense vectorial representation, which captures the trigger words more fuzzily. The contribution of the association features is less pronounced. The drug-problem (SemClass) and concept-concept (PMI) features have a small positive effect for TrP relations, with PMI working best (+0.5) for PP relations, where the coverage is the highest.

We now have a detailed look at the effect of the individual features. For this, we contrast the confusion matrix obtained from the base SegCNN with the confusion matrix of an extended model, where these matrices represent counts averaged over 20 runs. We obtain a new, contrasted matrix by subtracting the SegCNN matrix from that of the extended model, and display it as a heat map. An extension works well when the counts in the first diagonal are positive, and all the remaining counts are negative. In Figures 2a and 2b, we see an increase in correct classifications for semantic class features across all relation types, which speaks about the generality of this feature. The sensitivity for all relations has also increased (first column) as there are fewer true relations that remained unidentified. However, the counts of the less frequent relations (TrIP, TrNAP and TrWP) have shifted to incorrect relations (note the pale-red cells in the lower left corner of 2a). The improvements are the most obvious for the most frequent relations (TrAP and TeRP), with a clear increase in sensitivity, and a reduction in the number of unrelated (None) concepts classified as either TrAP or TeRP. The confusion matrix comparison for the problem-problem association (PMI) feature is shown in Figures 2c and 2d.[6] For TeP relations, we see that the addition of this feature type helps in reducing the number of hallucinated relations (first row), but at the expense of sensitivity—note that several relations are left unidentified (the counts in the TeCP and TeRP in the first column increased). A slight positive effect of PMI features can be seen for the PP relation, where the model becomes less prone to proclaim unrelated concepts as related (first row). Based on these figures, we can conclude that the PMI feature helps in deciding whether a pair of concepts should be linked with a relation or not, but does not have sufficient power to distinguish between different relations.

In conclusion, results show that the SegCNN model often misses, hallucinates or confuses relations, and that including semantic classes for relation triggers helps for different relation types.

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments. This research was carried out within the Accumulate strategic basic research project, funded by the government agency Flanders Innovation & Entrepreneurship (VLAIO) [grant number 150056].

References

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *ACL-IJCNLP*.

Dina Demner-Fushman, Emilia Apostolova, R Islamaj Dogan, et al. 2010. NLM's system description for the fourth i2b2/VA challenge. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2*.

Cyril Grouin, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deleger, Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard, Sophie Rosset, and Pierre Zweigenbaum. 2010. CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches. In *i2b2 Medication Extraction Challenge Workshop*.

Kai Hakala, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. 2016. Syntactic analyses and named entity recognition for pubmed and pubmed central — up-to-the-minute. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 102–107. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:.

Annervaz K. M., Somnath Basu Roy Chowdhury, and Ambedkar Dukkipati. 2018. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 313–322. Association for Computational Linguistics.

Yuan Luo, Yu Cheng, Özlem Uzuner, Peter Szolovits, and Justin Starren. 2017. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *Journal of the American Medical Informatics Association*, 25(1):93–98.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11).

Bryan Rink, Sanda Harabagiu, and Kirk Roberts. 2011. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600.

Sunil Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. 2016. Relation extraction from clinical texts using domain invariant convolutional neural network. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 206–215. Association for Computational Linguistics.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2017. CLAMP — a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

[6] We include the remaining TrP matrix and the matrices for the Drugbank model in Appendix.

	# documents	# TrP	# TeP	# PP
train	272 (64%)	2220	2233	1413
dev.	68 (16%)	587	485	325
test	86 (20%)	846	839	465

Table 2: Data statistics.

Maarten Van Gompel, Ko van der Sloot, and Antal van den Bosch. 2012. Ucto: Unicode Tokeniser. Technical report, Tilburg Centre for Cognition and Communication, Tilburg University and Radboud Centre for Language Studies, Radboud University Nijmegen.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2017. Clinical information extraction applications: A literature review. *Journal of biomedical informatics*.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2017. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in LSTMs for improving machine reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1436–1446.

A Supplemental Material

A.1 Experimental setup

Luo et al. (2017) used a part of the i2b2/VA dataset that is no longer available to those requesting the dataset. We therefore only have 170 documents for training and 256 documents for testing. Since our goal is to build an accurate relation extractor, we re-balance the dataset by increasing the size of the training corpus, reducing the size of the test set and creating a small development set. The sizes of the final splits are shown in Table 2. In all our experiments, we use the gold-standard concept annotations, and train one classifier per relation category.

Hyper-parameters We use the same set of hyper-parameters as Luo et al. (2017), except that we turn off the drop out on the final layer of the classifier network, which harmed the performance in our experiments on the development set. We also noticed that scaling of the added features positively affected the results, so we tuned the scaling factor as well.

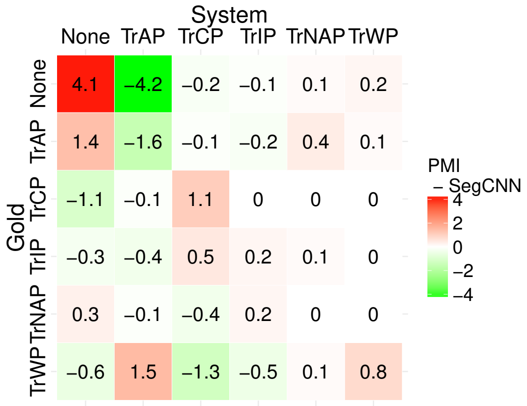
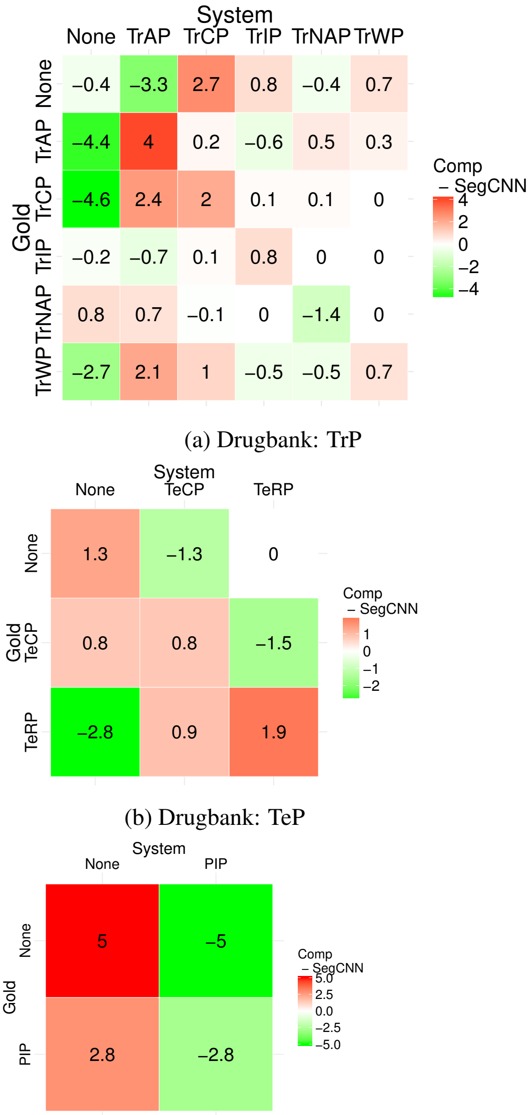


Figure 3: A comparison of counts between a base SegCNN and a model extended with PMI features, for different relation categories.

Embeddings We trained the word embeddings on a combination of PubMed abstracts, open-access PMC articles (Hakala et al., 2016) and MIMIC-III intensive care notes (Johnson et al., 2016), all segmented and tokenized, totaling around 9 billion tokens. We induce the embeddings using word2vec's CBOW model (Mikolov et al., 2013) and the default parameters, except for dimensionality, which we set to 200 for TrP relations, 500 for TeP and 400 for PP relations, as in Luo et al. (2017).

A.2 Supplementary results

The additional results from a contrastive confusion matrix analysis are shown in Figure 3 for the PMI extension, and in Figure 4 for the model with the added drug-treatment association feature.



(a) Drugbank: TrP

(b) Drugbank: TeP

(c) Drugbank: PP

Figure 4: A comparison of counts between a base SegCNN and a model extended with Drugbank features, for different relation categories.