Using Active Learning to Expand Training Data for Implicit Discourse Relation Recognition

Yang Xu Yu Hong* **Huibin Ruan Jianmin Yao Min Zhang Guodong Zhou**
Institute of Artificial Intelligence, Soochow University
School of Computer Science and Technology, Soochow University
No.1, Shizi ST, Suzhou, China, 215006
{andreaxu41, tianxianer, hbr416}@gmail.com
{jmyao}@szkj.gov.cn
{minzhang, gdzhou}@suda.edu.cn

Abstract

We tackle discourse-level relation recognition, a problem of determining semantic relations between text spans. Implicit relation recognition is challenging due to the lack of explicit relational clues. The increasingly popular neural network techniques have been proven effective for semantic encoding, whereby widely employed to boost semantic relation discrimination. However, learning to predict semantic relations at a deep level heavily relies on a great deal of training data, but the scale of the publicly available data in this field is limited. In this paper, we follow Rutherford and Xue (2015) to expand the training data set using the corpus of explicitly-related arguments, by arbitrarily dropping the overtly presented discourse connectives. On the basis, we carry out an experiment of sampling, in which a simple active learning approach is used, so as to take the informative instances for data expansion. The goal is to verify whether the selective use of external data not only reduces the time consumption of retraining but also ensures a better system performance. Using the expanded training data, we retrain a convolutional neural network (CNN) based classifer which is a simplified version of Qin et al. (2016)'s stacking gated relation recognizer. Experimental results show that expanding the training set with small-scale carefully-selected external data yields substantial performance gain, with the improvements of about 4% for accuracy and 3.6% for F-score. This allows a weak classifier to achieve a comparable performance against the state-of-the-art systems.

1 Introduction

Since the Penn Discourse Treebank of version 2.0 (PDTB) was released in 2008 (Prasad et al., 2008), there is a significant amount of research has been carried out on discourse-level relation recognition between a variety of text spans (namely, argument-argument relations). From a perspective of inclusion or omission of conjunctions, the study in this field has been directed toward two issues: recognizing explicit relations or implicit. Listed below are two pairs of arguments, where the arguments in 1) hold an explicit causal relation while those in 2) are implicitly related with a causal relation.

> 1) [She left the company]$_{Arg1}$ **because** [she would move to California]$_{Arg2}$.
>
> 2) [We have never seen the kitty since then]$_{Arg1}$. [John told us the kitty has been adopted]$_{Arg2}$.

In general, the explicit relations can be directly signaled by the conjunctions (also called connectives) which inherently exist, such as the conjunction "*because*" in 1). This allows a predictor to speculate relations by the word senses of conjunctions. Using conjunctions as relational predicates, the earlier study has achieved a prediction performance of no less than 93% for accuracy (Pitler and Nenkova, 2009). By contrast, the implicit relations like that in 2) are difficult to automatically recognize due to the lack of conjunctions.

We focus on the implicit relation recognition in this paper, and follow Rutherford and Xue (2015) to strengthen the current neural discourse-level relation classification by expanding the training data set. The explicit-to-implicit (Exp2Imp for short) relation transformation is used. In particular, we propose to introduce active learning into the data expansion process, with the aim to reduce redundancy and reinforce the use of informative instances. In our experiments, we cooperate active learning with a simple version of Exp2Imp transformation. Experimental results show that such a cooperation allows substantial improvements to be achieved for 4 main relation types in PDTB.

* Corresponding author

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 725–731
Brussels, Belgium, October 31 - November 4, 2018.

2 Related Work

Multi-class implicit relation recognition can be boiled down to a classification problem. This encorages the study of supervised classification at the earlier time (Pitler and Nenkova, 2009; Lin et al., 2009; Louis et al., 2010; Park and Cardie, 2012; Rutherford and Xue, 2014). Recently, the neural network based approaches become increasingly popular due to the capacity of deep semantic learning and understanding (Zhang et al., 2015; Qin et al., 2016; Chen et al., 2016; Qin et al., 2017; Liu and Li, 2016). However, a large amount of labeled data is urgently needed to train the models. (Rutherford and Xue, 2015; Braud and Denis, 2016; Liu et al., 2016; Wu et al., 2017).

The explicitly-related arguments in the corpus of PDTB has been sufficiently proven to be usable for creating implicitly-related arguments (Rutherford and Xue, 2015; Braud and Denis, 2016; Liu et al., 2016; Wu et al., 2017), only if the omission of inherent conjunctions will not distort the original semantic relations (Rutherford and Xue, 2015). Benefiting from the high-accuracy explicit relation recognition, a simple pattern, such as `Argument1`+*because*+`Argument2`, may enable the acquisition of countless explicitly-related arguments from texts. It makes it possible to cooperate with Rutherford and Xue (2015)'s Exp2Imp relation transformation for creating a tremendous number of labeled implicit relation instances.

However, it is inevitable to bring in the redundant information when using such a scale of unrefined artificial instances to directly expand the existing training data. This causes a time-consuming retraining process. By contrast, random sampling for retrenchment may leave informative instances out of the expanded data set.

Active learning (AL) is especially applicable to redundancy reduction. It is able to automatically select the most informative samples for use in a cycle of training. Nowadays, there have been a variety of AL models successfully used in different language processing tasks (Li and Guo, 2013; Guo and Wang, 2015; Yang et al., 2015; Zhang et al., 2017; Ramirez-Loaiza et al., 2017). This allows us to draw lessons from the experiences.

3 Informative Instances

In general, in the field of machine learning, an informative instance is defined as the one which has been classified with less confidence (i.e., higher uncertainty). Assume that $I_{r_j}(x_i, M)$ refers to the the level of uncertainty at which a classifier *M* determines an instance x_i as the member of the class r_j, thus x_i is informative only if it significantly increases the level of uncertainty:

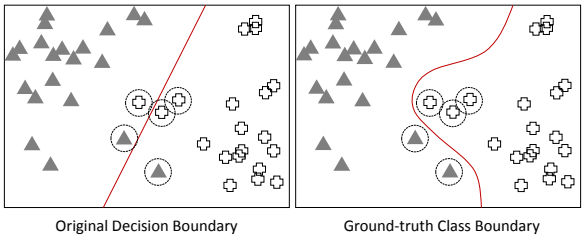

Figure 1: Examples of informative instances

$$x^* = \arg\max \sum_{r_j \in R} I_{r_j}(x_i; M) \quad (1)$$

The utilization of informative instances in data expansion for training has been proven effective in improving the fully-supervised classification models. For example, the instances marked with a dot-dashed circle in Figure 1 can be regarded to be informative. It is because a classifier may fail to deterministically distinguish between them. If using such instances as additional training data, we may retrain the classifier to revise the original decision boundary, and pursue the ground-truth.

4 Active Learning (AL)

AL is a kind of retraining mechanism by data expansion. Figure 2 shows the workflow, which primarily includes 4 steps:

- **Step 1**: in which a learning model is required to be trained on the previously labeled data.
- **Step 2**: where the well-trained model is used to classify the unlabeled external data.
- **Step 3**: relies on the classification results to evaluate informativeness over the unlabeled data. The informative instances will be eventually adopted for manual annotation.
- **Step 4**: adds the newly annotated data to the existing, and retrains the learning model.

It is noteworthy that AL is not an "once-for-all" deal. On the contrary, it needs to be carried out repeatedly and iteratively, until a predefined condition is met, such as the time at which the classification performance remains almost the same within

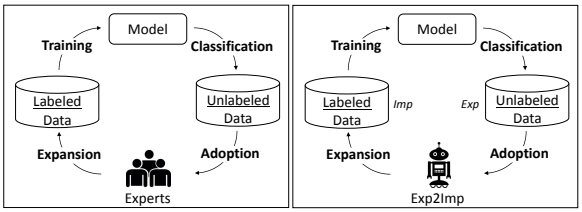

Figure 2: Workflows of ALs

a series of successive iterations, or termination after a fixed number of iterations.

5 Cooperating AL with Exp2Imp

We cooperate AL with Exp2Imp transformation. The workflow is also shown in Figure 2, where the labeled data set contains a certain number of implicitly-related argument pairs, while the unlabeled the explicitly-related.

In each iteration of AL, a classifier is trained on the labeled implicitly-related argument pairs, but it is driven to forcibly classify the explicitly-related argument pairs. On the basis, informativeness measurement is performed to sample the informative explicitly-related argument pairs. And then, instead of experts, the Exp2Imp transformation module serves as an annotator to mark the relations of the sampled argument pairs. Along with the automatically-annotated relations, such argument pairs are used as the counterfeit implicit relation instances. They are eventually added to the original labeled data set for expansion.

5.1 Informativeness Measurement

We employ an uncertainty sampling function (Zhu et al., 2008; Settles, 2010; Yang et al., 2015; Ramirez-Loaiza et al., 2017) to measure the informativeness:

$$Inf(x_i) = \sum_{r_j \in R} I_{r_j}(x_i; M) \quad (2)$$

$$= \sum_{r_j \in R} P(r_j|x_i) \log P(r_j|x_i) \quad (3)$$

where, for an instance x_i, the entropy of the predicted probabilities over all kinds of PDTB relation classes is used as the score of informativeness.

In order to reduce the computational complexity in practice, we sample informative instances in an iteration-independent batch-by-batch manner. In each iteration, a batch of instances in the unlabeled data set will be taken, only if their informativeness scores are higher than a constant threshold θ:

$$U' = \{x_i \mid Inf(x_i) > \theta, \forall x_i \in U\} \quad (4)$$

where, U is the unlabeled data set while U' consists of the potentially informative instances.

Learning rate	0.001	Filter size	$(2, 3, 5)$
Number of filter	512	Optimizer	Adam
Batch size	128	threshold θ	0.95

Table 1: Hyperparameter settings of CNN

5.2 Statistical Information based Exp2Imp

We implement a much simpler Exp2Imp transformation model than Rutherford and Xue (2015)'s work. Statistical information is used.

For an explicitly-related argument pair in U, we rely on the conjunction to determine the relation. For example, if the arguments are syntactically connected by the conjunction "*because*", their relation will be identified as `Contingency` (`Causality` plus `Condition`). We previously look up conjunctions in a small subset of U and analysis the ground-truth explicit relations they signal. If a conjunction invariably signals a single type of relation (e.g., because$\rightarrow$ `Contingency`), we preserve the one-to-one correspondence between the conjunction and the relation type. Else if a conjunction used to signal multiple-type relations, we align it solely with the relation it most frequently signals ever.

The Exp2Imp module first determines explicit relations based on conjunction-relation alignment, and then omits the conjunctions to create the counterfeit implicitly-related argument pairs.

5.3 CNN based Classification

We follow Qin et al. (2016) to use Siamese CNN for argument modeling and relation classification. The 300-dimensional word embeddings and 50-dimensional POS embeddings are used to represent the arguments. We also follow Mikolov et al. (2013) to pretrain the word embeddings and initialize the POS by random sampling in [-1,1]. Table 1 shows the hyperparameter settings.

The source codes of AL, Exp2Imp and Siamese CNN[1] to fully reproduce the experiments has been made publicly available, to which we attached the set of informative training data.

[1]https://github.com/AndreaXu0401/ALIDRC

Metrics		Baseline	Blender (U)	AL (U)
Tem.	P	61.54	66.67	81.82
	R	14.55	14.54	16.36
	F_1	23.53	23.88	**27.27**
Com.	P	38.74	72.72	82.50
	R	29.66	17.94	22.76
	F_1	33.59	28.73	**35.68**
Con.	P	53.15	45.64	59.00
	R	27.84	32.60	34.80
	F_1	36.54	38.03	**43.78**
Exp.	P	60.08	55.98	59.57
	R	83.09	79.92	88.48
	F_1	69.73	65.85	**71.21**
Accuracy		56.78	54.70	**60.63**
Macro F_1		40.85	39.12	**44.48**

Table 2: The four-way classification performance

6 Experimentation

We experiment on the PDTB v2.0 (Prasad et al., 2008). Sections 2-20 are used as the benchmark training set. They are also used as the labeled set in the AL process. Sections 21-22 are taken as the test set and sections 0-1 the development set.

The ground-truth explicitly-related argument pairs in the PDTB corpus are divided into two subsets: one consists of 450 instances which are used for aligning conjunctions and relation types, the other is consisted of 17,000 instances and used as the unlabeled data set U in the AL process.

6.1 Main Results

For the purpose of comparison, we expand the benchmark training data set (**Baseline**) in two different ways: **Blender** and **AL**. Blender combines the benchmark with U. Exp2Imp transformation is performed for the instances in U. Thus Blender mixes the true implicitly-related argument pairs and all the counterfeits in U. By contrast, AL samples informative counterfeits for expansion.

We train CNN on the benchmark and the expanded versions, and test the best-developed models for four-way classification among the relation types of `Expansion` (**Exp**), `Contingency` (**Con**), `Comparison` (**Com**) and `Temporality` (**Tem**). Table 2 shows the performance. It can be observed that simply adding all the counterfeits to the training data set negatively influences the learning process, causing a performance loss of about 1.73% for macro F_1 and 2.08% for accuracy. This may result from the fact that the relations of some counterfeits change to be uncertain due to the omission of inherent conjunctions. Such counterfeits probably mislead CNN during the learning process.

Systems	Accuracy	Macro F_1
IO (2015)	57.10	40.50
MNN (2016)	57.27	44.98
MTN (2016)	-	42.50
DSWE (2017)	58.85	44.84
MANN (2017)	57.39	**47.80**
Ours	**60.63**	44.48

Table 3: Comparison with the state of the art

On the contrary, AL obtains a substantial performance gain. Using about 15 percent of instances in U for expansion, AL improves CNN with 3.85% for accuracy and 3.63% for macro F_1. This may imply that the carefully-selected counterfeits by AL most probably have positive effects on the retraining. The uncertainty caused by arbitrary conjunction omission plays a counter-productive role, prompting CNN to pursue more precise classification boundaries in the AL process.

6.2 Discussion

6.2.1 Comparison to Expansion Methods

We compare the proposed method to the recently popular, all of which more or less expand the training data, so as to introduce comprehensive linguistic knowledge into the learning process:

- Intelligent Omission (**IO**) (Rutherford and Xue, 2015) consciously omits conjunctions in explicitly related argument pairs to create counterfeits. Herein, a model is developed to identify the counterfeits whose semantic relations keep unchanged after conjunction omission. Such samples are used for expansion.

- Multi-task Neural Network (**MNN**) (Liu et al., 2016). The training set is expanded with RST-DT (Carlson et al., 2003) and NYT (Sandhaus, 2008). The other two MNN based models include Wu et al. (2016)'s **MTN** and Lan et al. (2017)'s **MANN**. MANN is more sophisticated due to the use of attention mechanism, while MTN acts as a bilingual discourse analysis system. Both of them introduce external datasets in the training processes, NANT and BiSynData, respectively .

- **DSWE** based CNN model (Wu et al., 2017), where Discourse-Specific Word Embeddings (DSWE) are used. DSWE are pre-trained on the large English Gigaword corpus.

It can be found that our method achieves competitive performance, showing the best accuracy and almost comparable F score to those of most competitors. It is noteworthy that the neural network we use is much weaker than the sophisticated MNN, MTN and MANN. Besides, there are fewer instances used for expansion than that for training DSWE. This may imply that AL is cost-effective.

6.2.2 Informative Instances

We randomly sample a couple of informative instances by tracking the AL process, along with some less informative cases. They helps illustrate the intuition that informativeness verification benefits data cleaning, in the process of using counterfeits for distant supervision.

> 3) [He further said that it would study other alternatives]$_{Arg1}$ **omitted conjunction** [it hasn't yet made any proposals to shareholders]$_{Arg2}$.
>
> Conjunction: **however**
> Relation: `Comparison.Contrast`
>
> 4) [all of a sudden you are relegated to a paltry sum]$_{Arg1}$ **omitted conjunction** [you become a federal judge]$_{Arg2}$.
>
> Conjunction: **when**
> Relation:`Temporality.Synchrony`

The instances listed above are to some extent informative. When the conjunctions inherently connecting the arguments were pruned off, the semantic relations changed to be uncertain. For example, assume the conjunctions "*however*" and "*when*" are respectively replaced by "*therefore*" and "*although*", the arguments in 3) will appear to have a `Contingency.Causality` relation as well, and those in 4) seems to hold a `Contingency.Compromise` relation with a very reasonable possibility. It means that the relations the argument pair hold are ambiguous if there isn't any manually-edited explicit relational signal (e.g., conjunction). Nevertheless, from the other perspective, such instances are frankly useful for training a classifier. It is because:

- Learning to distinguish the relations of such arguments is challenging beyond other cases.
- A challenging task enables the training process to be more strict but effective.
- This makes it possible to learn the subtle differences among argument pairs which belong to different relation classes.

On the contrary, the instances listed below are far from informative. There are less alternative relations can be imaged to replace the original, even if the conjunctions "*although*" and "*because*" have been pruned off. Note that such instances are undoubtedly useful at the very beginning of the training process. At that time, they facilitate the initialization of fuzzy classification boundaries. However, they are most probably useless for calibrating the boundaries at the level of subtle difference.

> 5) [She was dreadful to her war-damaged husband]$_{Arg1}$ **omitted conjunction** [she was kind and playful to her children]$_{Arg2}$.
>
> Conjunction: **although**
> Relation: `Comparison.Contrast`
>
> 6) [The ad was devastating]$_{Arg1}$. **omitted conjunction** [it raised questions about Mr. Courter's credibility]$_{Arg2}$.
>
> Conjunction: **because**
> Relation: `Contingency.Cause`

7 Conclusion

This paper demonstrates the contributions of AL to the enhancement of implicit relation classification. Using an AL model, we successfully sample a batch of informative explicitly-related argument pairs. Following Exp2Imp strategy, we convert the arguments into implicitly-related cases. This helps expand training data for fully-supervised relation modeling. By retraining, we enable a weak CNN model to achieve competitive performance.

Active learning can be systematically cooperated with Rutherford and Xue (2015)'s intelligent omission. This will enable the sampling of both informative and reliable instances for data expansion. Besides, using the carefully expanded training data set, the sophisticated learning model like MNN may be further enhanced significantly.

Acknowledgments

This work was supported by the national Natural Science Foundation of China (NSFC) via Grant Nos. 61672368, 61672367, 61525205.

References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583. Association for Computational Linguistics.

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *EMNLP*, pages 203–213.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *ACL (1)*.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 547–558. Association for Computational Linguistics.

Husheng Guo and Wenjian Wang. 2015. An active learning-based svm multi-class classification model. *Pattern recognition*, 48(5):1577–1597.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308.

Xin Li and Yuhong Guo. 2013. Active learning with multi-label svm classification. In *IJCAI*, pages 1479–1485.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 343–351. Association for Computational Linguistics.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. *arXiv preprint arXiv:1609.06380*.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *AAAI*, pages 2750–2756.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 59–62. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2263–2270.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. *arXiv preprint arXiv:1704.00217*.

Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. 2017. Active learning: an empirical study of common baselines. *Data Mining and Knowledge Discovery*, 31(2):287–313.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *EACL*, volume 645, page 2014.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *HLT-NAACL*, pages 799–808.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Changxing Wu, Yidong Chen, Yanzhou Huang, et al. 2016. Bilingually-constrained synthetic data for implicit discourse relation recognition. In *Proceedings*

of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2306–2312.

Changxing Wu, Xiaodong Shi, Yidong Chen, Jinsong Su, and Boli Wang. 2017. Improving implicit discourse relation recognition with discourse-specific word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 269–274, Vancouver, Canada. Association for Computational Linguistics.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *EMNLP*, pages 2230–2235.

Ye Zhang, Matthew Lease, and Byron C Wallace. 2017. Active discriminative text representation learning. In *AAAI*, pages 3386–3392.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1137–1144. Association for Computational Linguistics.