Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models

Elena Tutubalina[1], **Artur Kadurin**[1] and **Zulfat Miftahutdinov**[1,2]
[1]Insilico Medicine Hong Kong, Pak Shek Kok, Hong Kong
[2]Kazan Federal University, Kazan, Russia
elena@insilico.com, artur@insilico.com, zulfatmi@gmail.com

Abstract

Linking of biomedical entity mentions to various terminologies of chemicals, diseases, genes, adverse drug reactions is a challenging task, often requiring non-syntactic interpretation. A large number of biomedical corpora and state-of-the-art models have been introduced in the past five years. However, there are no general guidelines regarding the evaluation of models on these corpora in single- and cross-terminology settings. In this work, we perform a comparative evaluation of various benchmarks and study the efficiency of state-of-the-art neural architectures based on Bidirectional Encoder Representations from Transformers (BERT) for linking of three entity types across three domains: research abstracts, drug labels, and user-generated texts on drug therapy in English. We have made the source code and results available at https://github.com/insilicomedicine/Fair-Evaluation-BERT.

1 Introduction

Aggregating knowledge about entities across different domains and corpora is critical for many information extraction (IE) applications. In biomedical research and healthcare, the entity linking problem is known as medical concept normalization (MCN). Medical concepts may have different types (e.g., drugs, diseases, or genes/proteins) and may be retrieved from different single-typed ontologies. Effective mapping of the same concepts across different ontologies (the MCN task) is the holy grail of modern medical NLP.

Most MCN methods meanwhile are evaluated on test sets of widely differing sizes and domains and a narrow subsample of concepts from specific terminology. Moreover, the reported results of neural networks vary substantially on different corpora, with, for example, accuracy ranging at least from 91% to 96% on research abstracts (Sung et al., 2020) and accuracy from 77% to 89% on social media texts (Miftahutdinov and Tutubalina, 2019).

Owing to their superior semantic learning capabilities, BERT (Devlin et al., 2019) and other neural architectures have been widely used in recent state-of-the-art (SOTA) models for the MCN task on research abstracts and social media texts (Leaman and Lu, 2016; Zhao et al., 2019; Li et al., 2017; Phan et al., 2019; Wright et al., 2019; Sung et al., 2020; Miftahutdinov and Tutubalina, 2019; Ji et al., 2020). These studies mostly share the same limitations regarding their evaluation strategy: models are usually trained and evaluated on entities of the same type from a single domain. Often, concept unique identifiers (CUIs) used in training are included in the test set. A recurring problem, which arises with supervised models, is how to reuse trained models for a different purpose; this requires coding to a specific terminology. In this work, we take the task a step further from existing research by exploring current benchmarks and cross-terminology transfer between entity mentions in research abstracts, drug labels, and user-generated texts.

We perform an extensive evaluation of five biomedical corpora manually annotated with concepts regarding diseases, chemicals, human genes, and adverse drug reactions (ADRs). We utilize two models:

Proceedings of the 28th International Conference on Computational Linguistics, pages 6710–6716
Barcelona, Spain (Online), December 8-13, 2020

	NCBI Disease	**BC5CDR Disease**	**BC5CDR Chem**	**BC2GN Gene**	**TAC 2017 ADR**	**SMM4H 2017 ADR**
domain	abstracts	abstracts	abstracts	abstracts	drug labels	tweets
entity type	disease	disease	chemicals	genes	ADRs	ADRs
terminology	MEDIC	MEDIC	CTD Chem	Entrez Gene	MedDRA	MedDRA
number of pre-processed entity mentions						
full corpus	6881	12850	15935	5712	13381	9150
avg. len in chars	20.37	14.88	11.27	8.35	17.28	11.69
% have numerals	5.74%	0.11%	7.32%	62.46%	1.62%	2.52%
train set	5134	4182	5203	2725	7038	6650
dev set	787	4244	5347	-	-	-
test set	960	4424	5385	2987	6343	2500
refined test	204 (21.2%)	657 (14.9%)	425 (7.9%)	985 (32.9%)	1,544 (24.3%)	831 (33.3%)
number of concepts						
train set $\|T_1\|$	668	968	922	556	1517	472
test set $\|T_2\|$	203	669	617	670	1323	254
refined test $\|T_3\|$	140	438	351	642	857	201
$\|T_1 \cap T_2\|$	136	457	368	55	867	218
$\|T_1 \cap T_3\|$	76	226	102	27	401	165

Table 1: Statistics of the datasets used in our experiments.

(i) a baseline that ranks concepts for a given mention by comparing biomedical BERT vectors (Lee et al., 2019) with the Euclidean distance; (ii) a supervised SOTA model BioSyn (Sung et al., 2020). The work reported here aims to advance SOTA models in biomedical concept normalization of entity mentions with a variety of entity types and differences in surface characteristics of mentions. In this work, we seek to answer the following research questions: **RQ1:** Do test sets of current benchmarks lead to an overestimation of performance? **RQ2:** How do surface characteristics of entity mentions affect the performance of the BERT-based baseline? **RQ3:** Does a model trained on one corpus work for the linking of entity mentions of another type or domain in the zero-shot setting?

2 Datasets and Resources

We use the following publicly available benchmarks with official train/dev/test splits. Descriptive statistics of these datasets are shown in Table 1.

NCBI Disease Corpus The NCBI Disease Corpus (Doğan et al., 2014) contains 793 PubMed abstracts with disease mentions and their concepts corresponding to the MEDIC dictionary (Davis et al., 2012). The NCBI corpus is the smallest (by the number of mentions), but the mentions have the longest average length and most of them are related to cancer and tumors. This MEDIC dictionary integrates concepts and synonyms from the Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2011) and the "Diseases" category of the National Library of Medicine's Medical Subject Headers (MeSH) (Coletti and Bleich, 2001). The "Diseases" category is very broad in MeSH; it includes conditions generally recognized as disease, abnormalities, injuries, poisoning, addiction, and pathological signs and symptoms. We use the MEDIC lexicon (v. July 6, 2012) that contains 11,915 CUIs and 71,923 synonyms.

BioCreative V CDR BioCreative V CDR (BC5CDR) (Li et al., 2016) introduces a task for the extraction of chemical-disease relations (CDR) from 1500 PubMed abstracts that contains annotations of both chemical/diseases. When dealing with chemicals, it is likely to see them expressed in the text exactly as they are seen in other abstracts: only 7.9% of mentions in the test set were unique or were not included in the train set. Disease and chemical mentions are linked to the MEDIC (Davis et al., 2012) and the Comparative Toxicogenomics Database (CTD) (Davis et al., 2019) dictionaries, respectively. We note

that CTD's chemical vocabulary is a modified subset of descriptors from the "Chemicals and Drugs" category and Supplementary Concept Records from MeSH. This category is very broad in MeSH; it includes therapeutic drugs, pure chemicals, and a variety of biological substances. The terms "drugs" and "chemicals" are often used interchangeably. We utilize the CTD chemical dictionary (v. November 4, 2019) that consists of 171,203 CUIs and 407,247 synonyms.

BioCreative II GN BioCreative II GN (BC2GN) (Morgan et al., 2008) contains PubMed abstracts with human gene and gene product mentions for gene normalization (GN) to Entrez Gene identifiers (Maglott et al., 2005). Gene mentions have the shortest average length and 62.46% contain numerals. To create the lexicon, we took the gene symbol, alias and description information for each gene identifier matched the following query on NCBI[1]: " 'Homo sapiens'[porgn] AND alive[prop]". It contains 61,646 CUIs and 277,944 synonyms.

TAC 2017 ADR TAC 2017 ADR (Roberts et al., 2017) proposes a challenge for the extraction of ADRs found in product labels (prescribing information or package inserts). ADRs are manually mapped into the MedDRA dictionary (Brown et al., 1999). In this study, we use MedDRA v19.0 which contains 24,033 CUIs and 77666 synonyms.

SMM4H 2017 ADR The Social Media Mining for Health (SMM4H) challenge (Sarker et al., 2018) presents a dataset with annotated ADR mentions linked to MedDRA. Tweets were collected using 250 generic and trade names for therapeutic drugs. Manually extracted ADR expressions were mapped to Preferred Terms (PTs) of the MedDRA dictionary. We use MedDRA v19.0 for this dataset. Out of the above-mentioned corpora, this corpus has the largest intersection of concepts between sets: 85% of concepts from the test set were present in the train set.

Preprocessing Similar to previous works (Leaman and Lu, 2016; Wright et al., 2019; Phan et al., 2019; Sung et al., 2020), we use several preprocessing steps. In particular, we adopt preprocessing scripts for datasets and dictionaries from the work (Sung et al., 2020). We use Ab3P (Sohn et al., 2008) to detect local abbreviations and replace each instance with the corresponding long form. We use heuristic rules (D'Souza and Ng, 2015) to split composite mentions into separate mentions (e.g., *non-familial breast and ovarian cancers* into *non-familial breast cancer* and *ovarian cancers*). Entity mentions from a training set are included in a corpus-specific dictionary. Finally, we process all characters to lowercase forms and remove punctuation for both mentions and synonyms.

2.1 Isolating train and test entity mentions

Given predefined splits of NCBI Disease, BC5CDR, and TAC 2017 ADR datasets, recent neural models achieve almost excellent accuracy averaging between 91% and 96% (Phan et al., 2019; Sung et al., 2020). Hence, one could view the MCN task on scientific texts as a largely solved task. After our analysis of datasets, we found out that approximately 80% entity mentions in the test set are textual duplicates of other entities in the test set or entities presented in train+dev sets. In order to obtain more realistic results, we present *refined* test sets without duplicates or exact overlaps. We note that some concepts appearing in the *refined* test set also appear in the respective training set (see $|T_1 \cap T_3|$ in Table 1).

In future work, we suggest that refined test sets can be split into two subsets, *stratified* and *zero-shot*. Here *Stratified* (Tutubalina et al., 2018) is intended to show how well models recognize known concepts with different surface forms of entity mentions. In contrast, the *zero-shot* setting shows how well models map mentions to novel concepts. Here we present a cross-terminology evaluation that is a more complicated version of zero-shot evaluation due to a shift in entity type and surface form mentions.

3 Models for Concept Normalization

We utilize two BERT-based models: (i) a baseline method based on the ranking of BioBERT representations, (ii) a SOTA model named BioSyn (Sung et al., 2020). We use BioBERT$_{base}$ v1.1 for both models that was pre-trained on PubMed abstracts (4.5B words in total) for 1M steps.

[1]https://www.ncbi.nlm.nih.gov/

BioBERT ranking This is a baseline model that used the BioBERT model for encoding mention and concept representations. Each entity mention or concept name is passed first through BioBERT (we use the average over all outputs of BERT) and then through a mean pooling layer to yield a fixed-sized vector. The inference task is then reduced to finding the closest concept name representation to entity mention representation in a common embedding space. We use the Euclidean distance as the distance metric. The nearest k concept names are chosen as top-k concepts for entities.

BioSyn BioSyn (Sung et al., 2020) is a recent SOTA model that utilizes the synonym marginalization technique and iterative candidate retrieval. The model uses two similarity functions based on sparse and dense representations, respectively. The sparse representation encodes the morphological information of given strings via TF-IDF, the dense representation encodes the semantic information gathered from BioBERT. BioSyn achieves SOTA results on NCBI, BC5CDR, TAC sets over previous works (Leaman and Lu, 2016; Wright et al., 2019; Phan et al., 2019). We have used the publicly available code provided by the authors at `https://github.com/dmis-lab/BioSyn` and reproduced the results successfully. We follow the default parameters of BioSyn as shown in (Sung et al., 2020): the number of top candidates $k = 20$, the mini-batch size is 16, the learning rate is 1e-5, the dense ratio for candidate retrieval is 0.5. We have trained BioSyn for 20 epochs for all datasets.

4 Single- and Cross-terminology Evaluation

We train BioSyn on the train/dev set of each corpus with a source dictionary, evaluating it on the respective test set (in-domain performance). For cross-domain evaluation, we assess models trained on *source* data on the test sets of all other corpora (i.e., the *target*). Specifically, both BioSyn and BioBERT ranking models retrieve the nearest concept name in a target dictionary for a given mention representation at inference time. We note that cross-terminology evaluation provides a challenging setup for developing supervised models, especially for linking to concepts not encountered during training (*zero-shot* concepts).

We evaluate this task in information retrieval (IR) scenario, where the goal is to find within a dictionary of concept names and their identifiers the top-k concepts for every entity mention in the texts. Let acc@k be 1 if the correct CUI is retrieved at rank k, otherwise 0. For composite entities, we define acc@k as 1 if every prediction for a single mention is correct. In particular, we use the top-1 accuracy as an evaluation metric, following previous works (Suominen et al., 2013; Pradhan et al., 2014; Wright et al., 2019; Phan et al., 2019; Sung et al., 2020).

Table 2 shows results on six sets where models are usually trained and evaluated on entities of the same type from a single domain. Table 3 compares the performance of BioSyn in single- and cross-terminology normalization tasks. Models were trained on the training set from a source dataset and evaluated on the target test set with different terminology.

To answer **RQ1**, we compare the results of models on official and *refined* test sets in Table 2. The significant decrease of averaged acc@1 from 91.8% to 76.7% for BioSyn and averaged acc@1 from 77.7% to 54.9% for BioBERT ranking highlights the great need for external evaluation datasets, where the same entity mentions will not be used for both training and testing. These observations also mean that there is room for improvement in the transferability of developed methods, that is, the ability to maintain performance for entirely unseen domains or entities.

According to Table 2, the following conclusions can be drawn to answer **RQ2**. First, the simple ranking of BioBERT representations achieves strong results on CDR Disease and Chemical sets. On two refined sets with larger mentions (NCBI, TAC) and the BC2GN corpus with mentions containing numerals, the difference between BioBERT ranking and BioSyn is significant (average decrease of 23.6%). Our qualitative analysis uncovered that BERT representations of mentions differing by one numeral (e.g., genes TP53 and TP63) are close in the latent space. As expected, results on SMM4H are significantly lower than on abstracts due to the gap between the language of lay public and medical professionals.

To answer **RQ3**, we compare performance differences in Tables 2 and 3. The models trained on NCBI, CDR Disease, BC2GN, and TAC data perform on par with the model trained on the CDR Chemical train set (approx. 74% acc@1), while the model trained on CDR Chemical showed a 6% drop on these subsets.

Model	NCBI Disease		BC5CDR Dis		BC5CDR Chem		BC2GN Gene		TAC ADR		SMM4H ADR	
	test	*refined*	test	*refined*	test	*refined*	test	*refined*	test	*refined*	test	*refined*
BioSyn	90.7	72.5	93.5	74.1	96.3	83.8	90.8	85.8	95.6	83.2	83.8	60.5
BioBERT ranking	83.9	47.5	91.3	65.1	94.7	79.3	74.7	68.4	87.8	54.7	33.9	14.3
Difference	-6.8	-25.0	-1.9	-7.7	-1.6	-4.5	-16.1	-17.4	-7.8	-28.5	-49.9	-46.2

Table 2: Single-terminology normalization results in terms of acc@1 on the official and *refined* test sets.

	Train set					
Test set	NCBI Dis	BC5CDR Dis	BC5CDR Chem	BC2GN Gene	TAC ADR	SMM4H ADR
NCBI Disease	72.5	67.6 (-4.9)	64.7 (-7.8)	67.2 (-5.4)	67.6 (-4.9)	48.5 (-24.0)
BC5CDR Dis	74.7 (+0.6)	74.1	73.4 (-0.8)	73.1 (-1.1)	74.9 (+0.8)	58.3 (-15.8)
BC5CDR Chem	82.4 (-1.4)	84.2 (+0.5)	83.8	82.6 (-1.2)	82.4 (-1.4)	73.9 (-9.9)
BC2GN Gene	83.1 (-2.6)	81.7 (-4.1)	83.7 (-2.1)	85.8	82.6 (-3.1)	73.2 (-12.6)
TAC ADR	74.3 (-8.9)	77.5 (-5.7)	70.1 (-13.0)	69.9 (-13.3)	83.2	51.5 (-31.7)
SMM4H ADR	27.3 (-33.2)	35.6 (-24.9)	24.8 (-35.7)	21.9 (-38.6)	30.1 (-30.4)	60.5

Table 3: Comparison of BioSyn for single- and cross-terminology MCN on *refined* test sets by accuracy@1. In-domain results are on the diagonals (with a dark gray background). Other cells contain results of a given model and differences in results between that model and the in-domain model in parentheses (by row). Light gray cells show cross-terminology experiments.

BioSyn trained on SMM4H achieves lower results on abstracts and drug labels than simple BioBERT ranking, while all supervised models performed better on SMM4H data than the BioBERT ranking.

5 Conclusion

We have presented the first comparative evaluation of medical concept normalization (MCN) datasets, studying the NCBI Disease, BC5CDR Disease & Chemical, BC2GN Gene, TAC 2017 ADR, and SMM4H 2017 ADR corpora. We perform an extensive evaluation of two BERT-based models on six datasets in two setups: with official train/test splits and with the proposed test sets that represent *refined* samples of entity mentions. Our evaluation shows great divergence in performance between these two test sets, finding an average accuracy difference of 15% for the state-of-the-art model BioSyn. We also performed a quantitative evaluation of BioSyn in the cross-terminology MCN task where models were trained and evaluated on entity mentions of various types with concepts from different terminologies. Knowledge transfer can be effective between diseases, chemicals, and genes with an average drop of 2.53% accuracy in the performance on NCBI, BC5CDR, and BC2GN sets. For TAC and SMM4H sets with ADRs from drug labels and social media, BioSyn models trained on four other corpora show a substantial decrease in performance (-10.2% and -33.1% accuracy, respectively) compared to in-domain trained models. To our surprise, these models still outperformed the straightforward ranking baseline on BioBERT representations. We believe that *refined* datasets with cross-terminology evaluation can serve as a step toward reliable and large-scale evaluation of biomedical IE models.

We foresee three directions for future work. First, a promising research direction is the multilingual evaluation of MCN models. Second, in some cases current models choose a broader concept that is in a parent-child relationship with the correct concept; here future research may focus on the encoding of the concept hierarchy. Third, the use of local and global contexts of entity mentions remains to be explored.

Acknowledgements

Work on social media data was carried out by Z.M. and supported by RFBR, project no. 19-37-90074. Work on problem definition was carried out by E.T. and supported by RSF, project no. 18-11-00284. We thank the anonymous referees and Sergey Nikolenko for insightful comments that helped improve the paper.

References

Joanna Amberger, Carol Bocchini, and Ada Hamosh. 2011. A new face and new challenges for online mendelian inheritance in man (omim®). *Human mutation*, 32(5):564–567.

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

Margaret H Coletti and Howard L Bleich. 2001. Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8(4):317–323.

Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. 2012. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, 2012.

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. 2019. The comparative toxicogenomics database: update 2019. *Nucleic acids research*, 47(D1):D948–D954.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302.

Zongcheng Ji, Qiang Wei, and Hua Xu. 2020. Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings*, 2020:269.

Robert Leaman and Zhiyong Lu. 2016. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. Cnn-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18(11):79–86.

Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. 2005. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58.

Zulfat Miftahutdinov and Elena Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 393–399.

Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, et al. 2008. Overview of biocreative ii gene normalization. *Genome biology*, 9(S2):S3.

Minh C Phan, Aixin Sun, and Yi Tay. 2019. Robust representation learning of biomedical names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285.

Sameer Pradhan, Noémie Elhadad, Wendy W Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *SemEval@ COLING*, pages 54–62.

Kirk Roberts, Dina Demner-Fushman, and Joseph M Tonning. 2017. Overview of the tac 2017 adverse reaction extraction from drug labels track. In *TAC*.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9(1):402.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. *arXiv preprint arXiv:2005.00239*.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics*, 84:93–102.

Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. 2019. Normco: Deep disease normalization for biomedical knowledge base construction. In *Automated Knowledge Base Construction*.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 817–824.