Sieve-Based Entity Linking for the Biomedical Domain

Jennifer D'Souza and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{jld082000,vince}@hlt.utdallas.edu

Abstract

We examine a key task in biomedical text processing, normalization of disorder mentions. We present a multi-pass sieve approach to this task, which has the advantage of simplicity and modularity. Our approach is evaluated on two datasets, one comprising clinical reports and the other comprising biomedical abstracts, achieving state-of-the-art results.

1 Introduction

Entity linking is the task of mapping an entity mention in a text document to an entity in a knowledge base. This task is challenging because (1) the same word or phrase can be used to refer to different entities, and (2) the same entity can be referred to by different words or phrases. In the biomedical text processing community, the task is more commonly known as normalization, where the goal is to map a word or phrase in a document to a unique concept in an ontology (based on the description of that concept in the ontology) after disambiguating potential ambiguous surface words or phrases. Unlike in the news domain, in the biomedical domain it is rare for the same word or phrase to refer to multiple different concepts. However, different words or phrases often refer to the same concept. Given that mentions in biomedical text are relatively unambiguous, normalizing them involves addressing primarily the second challenge mentioned above.

The goal of this paper is to advance the state of the art in normalizing *disorder mentions* in documents from two genres, clinical reports and biomedical abstracts. For example, given the disorder mention *swelling of abdomen*, a normalization system should map it to the concept in the ontology associated with the term *abdominal distention*. Not all disorder mentions can be mapped to a given ontology, however. The reason is that the ontology may not include all of the possible concepts. Hence, determining whether a disorder mention can be mapped to a concept in the given ontology is part of the normalization task. Note that disorders have been the target of many research initiatives in the biomedical domain, as one of its major goals is to alleviate health disorders.

	ShARe (Clinical Reports)		**NCBI (Biomedical Abstracts)**	
	Train	**Test**	**Train**	**Test**
Documents	199	99	692	100
Disorder mentions	5816	5351	5921	964
Mentions w/ ID	4178	3615	5921	964
ID-less mentions	1638	1736	0	0

Table 1: Corpus statistics.

Our contributions are three-fold. First, we propose a simpler and more modular approach to normalization than existing approaches: a multi-pass sieve approach. Second, our system achieves state-of-the-art results on datasets from two genres, clinical reports and biomedical abstracts. To our knowledge, we are the first to present normalization results on two genres. Finally, to facilitate comparison with future work on this task, we release the source code of our system.[1]

2 Corpora

We evaluate our system on two standard corpora (see Table 1 for their statistics):

The **ShARe/CLEF eHealth Challenge corpus** (Pradhan et al., 2013) contains 298 de-identified clinical reports from US intensive care partitioned into 199 reports for training and 99 reports for testing. In each report, each disorder mention is manually annotated with either the unique identifier of the concept in the reference ontology to which it refers, or "CUI-less" if it cannot be mapped to any

[1]The code is available from http://www.hlt.utdallas.edu/~jld082000/normalization/.

(1) C0000731||swollen abdomen|abdominal distension|abdomen distended|abdominal distention|abdominal swelling

(2) D008288||Malaria|Fever, Marsh|Fever, Remittent|Infection, Plasmodium|MALS|Plasmodium Infection|Remittent Fever

Figure 1: Example concepts in the ontologies. The first one is taken from SNOMED-CT and the second one is taken from MEDIC ontologies. In each concept, only its ID and the list of terms associated with it are shown.

concept in the reference ontology. The reference ontology used is the SNOMED-CT resource of the UMLS Metathesaurus (Campbell et al., 1998), which contains 128,430 disorder concepts.

The **NCBI disease corpus** (Doğan et al., 2014) contains 793 biomedical abstracts partitioned into 693 abstracts for training and development and 100 abstracts for testing. Similar to the ShARe corpus, a disorder mention in each abstract is manually annotated with the identifier of the concept in the reference ontology to which it refers. The reference ontology used is the MEDIC lexicon (Davis et al., 2012), which contains 11,915 disorder concepts. Unlike in the ShARe corpus, in NCBI only those disorder mentions that can be mapped to a concept in MEDIC are annotated. As a result, all the annotated disorder mentions in the NCBI corpus have a concept identifier. Unlike in ShARe, in NCBI there exist *composite* disorder mentions, each of which is composed of more than one disorder mention. A composite disorder mention is annotated with the set of the concept identifiers associated with its constituent mentions.

We note that each concept in the two ontologies (the UMLS Metathesaurus and MEDIC) is not only identified by a concept ID, but also associated with a number of attributes, such as the list of *terms* commonly used to refer to the concept, the preferred term used to refer to the concept, and its definition. In our approach, we use only the list of terms associated with each concept ID in the normalization process. Figure 1 shows two example concepts taken from these two ontologies.

3 A Multi-Pass Approach to Normalization

Despite the simplicity and modularity of the multi-pass sieve approach and its successful application to coreference resolution (Raghunathan et al., 2010), it has not been extensively applied to other NLP tasks. In this section, we investigate its application to normalization.

3.1 Overview of the Sieve Approach

A sieve is composed of one or more heuristic rules. In the context of normalization, each rule *normalizes* (i.e., assigns a concept ID to) a disorder mention in a document. Sieves are ordered by their precision, with the most precise sieve appearing first. To normalize a set of disorder mentions in a document, the normalizer makes multiple passes over them: in the i-th pass, it uses only the rules in the i-th sieve to normalize a mention. If the i-th sieve cannot normalize a mention *unambiguously* (i.e., the sieve normalizes it to more than one concept in the ontology), the sieve will leave it unnormalized. When a mention is normalized, it is added to the list of terms associated with the ontology concept to which it is normalized. This way, later sieves can exploit the normalization decisions made in earlier sieves. Note that a normalization decision made earlier cannot be overridden later.

3.2 Normalization Sieves

In this subsection, we describe the ten sieves we designed for normalization. For convenience, we use the word *concept* to refer to a concept in the ontology, and we say that a disorder mention has an exact match with a concept if it has an exact match with one of the *terms* associated with it.

Sieve 1: Exact Match. This sieve normalizes a disorder mention m to a concept c if m has an exact match with c.

Sieve 2: Abbreviation Expansion. This sieve first expands all abbreviated disorder mentions using Schwartz and Hearst's (2003) algorithm and the Wikipedia list of disorder abbreviations.[2] Then, it normalizes a disorder mention m to a concept c if the unabbreviated version of m has an exact match with c.

For each unnormalized mention, we pass both its original form and its new (i.e., unabbreviated) form, if applicable, to the next sieve. As we will see, we keep expanding the set of possible forms of an unnormalized mention in each sieve. Whenever a subsequent sieve processes an unnormalized mention, we mean that it processes each form of the mention created by the preceding sieves.

Sieve 3: Subject $\Leftrightarrow$ Object Conversion. This

[2] http://en.wikipedia.org/wiki/List_of_abbreviations_for_diseases_and_disorders

sieve normalizes a mention to a concept c if any of its new forms has an exact match with c. New forms of a mention m are created from its original and unabbreviated forms by: (1) replacing any preposition(s) in m with other prepositions (e.g., "changes on ekg" converted to "changes in ekg"); (2) dropping a preposition from m and swapping the substrings surrounding it (e.g., "changes on ekg" converted to "ekg changes"); (3) bringing the last token to the front, inserting a preposition as the second token, and shifting the remaining tokens to the right by two (e.g., "mental status alteration" converted to "alteration in mental status"); and (4) moving the first token to the end, inserting a preposition as the second to last token, and shifting the remaining tokens to the left by two (e.g., "leg cellulitis" converted to "cellulitis of leg"). As in Sieve 2, for each unnormalized mention in this and all subsequent sieves, both its original and new forms are passed to the next sieve.

Sieve 4: Numbers Replacement. For a disorder mention containing numbers between one to ten, new forms are produced by replacing each number in the mention with other forms of the same number. Specifically, we consider the numeral, roman numeral, cardinal, and multiplicative forms of a number for replacement. For example, three new forms will be created for "three vessel disease": {"3 vessel disease", "iii vessel disease", and "triple vessel disease"}. This sieve normalizes a mention m to a concept c if one of the new forms of m has an exact match with c.

Sieve 5: Hyphenation. A disorder mention undergoes either hyphenation (if it is not already hyphenated) or dehyphenation (if it is currently hyphenated). Hyphenation proceeds as follows: the consecutive tokens of a mention are hyphenated one pair at a time to generate a list of hyphenated forms (e.g., "ventilator associated pneumonia" becomes {"ventilator-associated pneumonia", "ventilator associated-pneumonia"}). Dehyphenation proceeds as follows: the hyphens in a mention are removed one at a time to generate a list of dehyphenated forms (e.g., "saethre-chotzen syndrome" becomes "saethre chotzen syndrome"). This sieve normalizes a mention m to a concept c if one of the new forms of m has an exact match with c.

Sieve 6: Suffixation. Disorder mentions satisfying suffixation patterns manually observed in the training data are suffixated. For example, "infectious source" becomes "source of infectious" in Sieve 3, which then becomes "source of infection" in this sieve. This sieve normalizes a mention m to a concept c if the suffixated form of m has an exact match with c.

Sieve 7: Disorder Synonyms Replacement. For mentions containing a disorder term, new forms are created by replacing the disorder term with its synonyms.[3] For example, "presyncopal events" becomes {"presyncopal disorders", "presyncopal episodes", etc.}. In addition, one more form is created by dropping the disorder modifier term (e.g., "iron-overload disease" becomes "iron overload disease" in Sieve 5, which becomes "iron overload" in this sieve). For mentions that do not contain a disorder term, new forms are created by appending the disorder synonyms to the mention. E.g., "crohns" becomes {"crohns disease", "crohns disorder", etc.}. This sieve normalizes a mention m to a concept c if any of the new forms of m has an exact match with c.

Sieve 8: Stemming. Each disorder mention is stemmed using the Porter (1980) stemmer, and the stemmed form is checked for normalization by exact match with the stemmed concept terms.

Sieve 9: Composite Disorder Mentions/Terms. A disorder mention/concept term is *composite* if it contains more than one concept term. Note that composite concept terms only appear in the UMLS ontology (i.e., the ontology for the ShARe dataset), and composite disorder mentions only appear in the NCBI corpus. Hence, different rules are used to handle the two datasets in this sieve. In the ShARe corpus, we first split each composite term associated with each concept in the UMLS ontology (e.g., "common eye and/or eyelid symptom") into separate phrases (e.g., {"common eye symptom", "common eyelid symptom"}), so each concept may now be associated with additional terms (i.e., the split terms). This sieve then normalizes a mention to a concept c if it has an exact match with c. In the NCBI corpus, we consider each disorder mention containing "and", "or", or "/" as composite, and split each such composite mention into its constituent mentions (e.g., "pineal and retinal tumors" is split into {"pineal tumors", "retinal tumors"}). This sieve then normalizes a composite mention m to a concept c as follows. First, it normalizes each of its split mentions to a concept c if the split mention has an exact match

[3] A list of the disorder word synonyms is manually created by inspection of the training data.

with c. The normalized form of m will be the union of the concepts to which each of its split mentions is normalized.[4]

Sieve 10: Partial Match. Owing to the differences in the ontologies used for the two domains, the partial match rules for the ShARe corpus are different from those for the NCBI corpus. In ShARe, a mention m is normalized to a concept c if one of the following ordered set of rules is applicable: (1) m has more than three tokens and has an exact match with c after dropping its first token or its second to last token; (2) c has a term with exactly three tokens and m has an exact match with this term after dropping its first or middle token; and (3) all of the tokens in m appear in one of the terms in c and vice versa. In NCBI, a mention is normalized to the concept with which it shares the most tokens. In the case of ties, the concept with the fewest tokens is preferred.

Finally, the disorder mentions not normalized in any of the sieves are classified as "CUI-less".

4 Related Work

In this section, we focus on discussing the two systems that have achieved the best results reported to date on our two evaluation corpora. We also discuss a state-of-the-art open-domain entity-linking system whose underlying approach is similar in spirit to ours.

DNorm (Leaman et al., 2013), which adopts a pairwise learning-to-rank approach, achieves the best normalization result on NCBI. The inputs to their system are linear vectors of paired query mentions and candidate concept terms, where the linear vectors are obtained from a tf-idf vector space representation of all unique tokens from the training disorder mentions and the candidate concept terms. Among all the candidate concepts that a given query disorder mention is paired with, the system normalizes the query mention to the highest ranked candidate. Similarity scores for ranking the candidates are computed by multiplying the linear tf-idf vectors of the paired query-candidate mentions and a learned weight matrix. The weight matrix represents all possible pairs of the unique tokens used to create the tf-idf vector. At the beginning of the learning phase, the weight matrix is initialized as an identity matrix. The matrix weights are then iteratively adjusted by stochastic gradient descent for all the concept terms, their matched training data mentions, and their mismatched training data mentions. After convergence, the weight matrix is then employed in the scoring function to normalize the test disorder mentions.

Ghiasvand and Kate's (Ghiasvand and Kate, 2014) system has produced the best results to date on ShARe. It first generates variations of a given disorder word/phrase based on a set of learned edit distance patterns for converting one word/phrase to another, and then attempts to normalize these query phrase variations by performing exact match with a training disorder mention or a concept term.

Rao et al.'s (2013) open-domain entity-linking system adopts an approach that is similar in spirit to ours. It links organizations, geo-political entities, and persons to the entities in a Wikipedia-derived knowledge base, utilizing heuristics for matching mention strings with candidate concept phrases. While they adopt a learning-based approach where the outcomes of the heuristics are encoded as features for training a ranker, their heuristics, like ours, employ syntactic transformations of the mention strings.

5 Evaluation

In this section, we evaluate our multi-pass sieve approach to normalization of disorder mentions. Results on normalizing gold disorder mentions are shown in Table 2, where performance is reported in terms of accuracy (i.e., the percentage of gold disorder mentions correctly normalized).

Row 1 shows the baseline results, which are the best results reported to date on the ShARe and NCBI datasets by Leaman et al. (2013) and Ghiasvand and Kate (2014), respectively. As we can see, the baselines achieve accuracies of 89.5 and 82.2 on ShARe and NCBI, respectively.

The subsequent rows show the results of our approach when our ten sieves are added incrementally. In other words, each row shows the results obtained after adding a sieve to the sieves in the previous rows. Our best system results, highlighted in bold in Table 2, are obtained when all our ten sieves are employed. These results are significantly better than the baseline results (paired t-tests, $p < 0.05$).

To better understand the usefulness of each sieve, we apply paired t-tests on the results in adjacent rows. We find that among the ten sieves,

[4] Note that a composite mention in NCBI may be associated with multiple concepts in the ontology.

	ShARe	NCBI
BASELINE	89.5	82.2
OUR SYSTEM		
Sieve 1 (Exact Match)	84.04	69.71
+ Sieve 2 (Abbrev.)	86.13	74.17
+ Sieve 3 (Subj/Obj)	86.40	74.27
+ Sieve 4 (Numbers)	86.45	75.00
+ Sieve 5 (Hyphen)	86.62	75.21
+ Sieve 6 (Affix)	88.11	75.62
+ Sieve 7 (Synonyms)	88.45	76.56
+ Sieve 8 (Stemming)	90.47	77.70
+ Sieve 9 (Composite)	90.53	78.00
+ Sieve 10 (Partial)	**90.75**	**84.65**

Table 2: Normalization accuracies on the test data from the ShARe corpus and the NCBI corpus.

Sieve 2 improves the results on both datasets at the lowest significance level ($p < 0.02$), while Sieves 6, 7, 8, and 10 improve results on both datasets at a slightly higher significance level ($p < 0.05$). Among the remaining four sieves (3, 4, 5, 9), Sieve 3 improves results only on the clinical reports ($p < 0.04$), Sieve 4 improves results only on the biomedical abstracts dataset ($p < 0.02$), and Sieves 5 and 9 do not have any significant impact on either dataset ($p > 0.05$). The last finding can be ascribed to the low proportions of hyphenated (Sieve 5) and composite (Sieve 9) disorder mentions found in the test datasets. After removing Sieves 5 and 9, accuracies drop insignificantly ($p > 0.05$) by 0.3% and 1.14% on the clinical reports and biomedical abstracts, respectively.

6 Error Analysis

In this section, we discuss the two major types of error made by our system.

Failure to unambiguously resolve a mention. Errors due to ambiguous normalizations where a disorder mention is mapped to more than one concept in the Partial Match sieve comprise 11–13% of the errors made by our system. For example, "aspiration" can be mapped to "pulmonary aspiration" and "aspiration pneumonia", and "growth retardation" can be mapped to "fetal growth retardation" and "mental and growth retardation with amblyopia". This ambiguity typically arises when the disorder mention under consideration is anaphoric, referring to a previously mentioned entity in the associated text. In this case, context can be used to disambiguate the mention. Specifically, a coreference resolver can first be used to identify the coreference chain to which the ambiguous mention belongs, and then the ambiguous mention can be normalized by normalizing its coreferent yet unambiguous counterparts instead.

Normalization beyond syntactic transformations. This type of error accounts for about 64–71% of the errors made by our system. It occurs when a disorder mention's string is so lexically dissimilar with the concept terms that none of our heuristics can syntactically transform it into any of them. For example, using our heuristics, "bleeding vessel" cannot be matched with any of the terms representing its associated concept, such as "vascular hemorrhage", "rupture of blood vessel", and "hemorrhage of blood vessel". Similarly, "dominantly inherited neurodegeneration" cannot be matched with any of the terms representing its associated concept, such as "hereditary neurodegenerative disease". In this case, additional information beyond a disorder mention's string and the concept terms is needed to normalize the mention. For example, one can exploit the contexts surrounding the mentions in the training set. Specifically, given a test disorder mention, one may first identify a disorder mention in the training set that is "sufficiently" similar to it based on context, and then normalize it to the concept that the training disorder mention is normalized to. Another possibility is to exploit additional knowledge bases such as Wikipedia. Specifically, one can query Wikipedia for the test mention's string, then employ the titles of the retrieved pages as alternate mention names.

7 Conclusion

We have presented a multi-pass sieve approach to the under-studied task of normalizing disorder mentions in the biomedical domain. When normalizing the gold disorder mentions in the ShARe and NCBI corpora, our approach achieved accuracies of 90.75 and 84.65, respectively, which are the best results reported to date on these corpora. Above all, to facilitate comparison with future work, we released the source code of our normalization system.

Acknowledgments

We thank the three anonymous reviewers for their detailed and insightful comments on an earlier draft of this paper. This work was supported in part by NSF Grants IIS-1147644 and IIS-1219142.

References

Keith E. Campbell, Diane E. Oliver, and Edward H. Shortliffe. 1998. The Unified Medical Language System: Towards a collaborative approach for solving terminologic problems. *Journal of the American Medical Informatics Assocication*, 5(1):12–16.

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:bar065.

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.

Omid Ghiasvand and Rohit Kate. 2014. UWM: Disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 828–832.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics*, pages 2909–2917.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Sameer Pradhan, Noemie Elhadad, B South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, W Chapman, and Guergana Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. *Online Working Notes of CLEF*, 230.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multi-lingual Information Extraction and Summarization*, pages 93–115.

Ariel Schwartz and Marti Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the 8th Pacific Symposium on Biocomputing*, pages 451–462.