SimConcept: A Hybrid Approach for Simplifying Composite Named Entities in Biomedicine

Chih-Hsuan Wei
8600 Rockville Pike, National Center for Biotechnology Information (NCBI), for Biotechnology Information (NCBI), for Biotechnology Information (NCBI), Bethesda, Maryland, USA, 20894
Chih-Hsuan.Wei@nih.gov

Robert Leaman
8600 Rockville Pike, National Center for Biotechnology Information (NCBI), Bethesda, Maryland, USA, 20894
Robert.Leaman@nih.gov

Zhiyong Lu
8600 Rockville Pike, National Center for Biotechnology Information (NCBI), Bethesda, Maryland, USA, 20894
Zhiyong.Lu@nih.gov

ABSTRACT
Many text-mining studies have focused on the issue of named entity recognition and normalization, especially in the field of biomedical natural language processing. However, entity recognition is a complicated and difficult task in biomedical text. One particular challenge is to identify and resolve composite named entities, where a single span refers to more than one concept (e.g., BRCA1/2). Most bioconcept recognition and normalization studies have either ignored this issue, used simple ad-hoc rules, or only handled coordination ellipsis, which is only one of the many types of composite mentions studied in this work. No systematic methods for simplifying composite mentions have been previously reported, making a robust approach greatly needed. To this end, we propose a hybrid approach by integrating a machine learning model with a pattern identification strategy to identify the antecedent and conjuncts regions of a concept mention, and then reassemble the composite mention using those identified regions. Our method, which we have named SimConcept, is the first method to systematically handle most types of composite mentions. Our method achieves high performance in identifying and resolving composite mentions for three fundamental biological entities: genes (89.29% in F-measure), diseases (85.52% in F-measure) and chemicals (84.04% in F-measure). Furthermore, our results show that, using our SimConcept method can subsequently help improve the performance of gene and disease concept recognition and normalization.

Categories and Subject Descriptors
I.2.7 [ARTIFICIAL INTELLIGENCE]: Natural Language Processing – Language parsing and understanding, Speech recognition and synthesis, Text analysis.

General Terms
Algorithms

Keywords
Mention simplification, conditional random field, natural language processing, name entity recognition, name entity normalization

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

BCB '14, September 20 - 23 2014, Newport Beach, CA, USA
ACM 978-1-4503-2649-4/14/09...
http://dx.doi.org/10.1145/2649387.2649420

1. INTRODUCTION
In biomedical text mining, many studies have focused on automatically extracting relevant information from published literature [1]. The relevant information is commonly focused on a specific topic, such as protein-protein interactions [2, 3]; protein transport and localization [4-6]; drug-disease associations [7-9], or gene function extraction [10, 11]. Most of the common retrieval methods apply natural language processing or machine learning to identify relations in text. One crucial step towards this goal is automatically recognizing bioconcept mentions (e.g., gene/protein) – the task of named entity recognition (NER) – and mapping the bioconcept to a specific database identifier (e.g., NCBI EntrezGene) – the task of normalization. Many international biomedical text mining competitions (e.g., BioCreative) have therefore focused on these tasks [12-16]. Genes, diseases and chemicals are particularly notable for not only being important concepts, but also being the most popular concepts in biomedical literature search [17]. Most normalization studies of different concepts face two challenges: term variation and ambiguity [18-22]. Many previous studies have defined individual strategies (e.g., machine learning, statistical inference and rule-based methods) to deal with these two issues. However, a particular type of error which has not been handled well is composite mentions, where a single span refers to more than one concept (e.g., “SMADs 1, 5, and 8”). We observe that in our datasets, approximately 10% of gene, disease and chemical mentions are composite mentions, hence it is important to handle them properly. This study presents a new method for bio-concept mention simplification in a systematical fashion.

Most previous related studies have focused on text (including document/paragraph [23-28] and sentence [29-35] levels) simplification. An early text simplification paper was introduced by Chandrasekar et al. [26, 27], which used a syntactic-parsing method (i.e., parse trees) to improve the performance of text simplification. In the past several years, most of the studies have focused on sentence simplification using three methodologies: lexical, syntactic and discourse simplification [29]. Leroy et al. [23] designed an algorithm that uses term familiarity to identify difficult text and select easier alternatives from lexical resources. Peng et al. [30] developed Simp which is an alternative method by using shallow parsing to detect the various simplification structures in text in linear time. Most of the studies on text simplification have focused on documents or sentences. The few studies which have concentrated on mention simplification have only addressed coordination ellipsis. Buyko et al., [36] developed a CRF-based method with three states: conjunction, conjuncts, and ellipsis antecedent. For an example “human and mouse cells”, “human” and “mouse” are conjuncts, “and” is conjunction, and “cells” is ellipsis antecedent. In their evaluation on GENIA [37] corpus, they obtained 86% accuracy on elliptical entity
expression evaluation. Due to the lower performance of this method on complex ellipsis (e.g., “recombinant human nm23-H1, -H2, mouse nm23-M1, and -M2”), Chae et al., [38] developed a pattern based method and used lexicons to identify the region of each component (i.e., conjunction, conjuncts, and ellipsis antecedent) for each mention. However, these previous studies have focused on only one type of composite mention: mentions with coordination ellipsis. Due to the demand for improving bioconcept identification and lack of related studies, we propose a hybrid method to handle most of the composite mention types. This study proposes a CRF-based model and a pattern identification strategy to detect the antecedent, suffix/strain and conjunction from a composite mention, which refers to more than one concept. Instead of building heuristic rules to match all possible variants, we propose a machine learning approach to simplify the mention. For example, after identifying all concepts from “SMADs 1, 5, and 8”, three individual concepts “SMAD1”, “SMAD5” and “SMAD8” will be generated. Some mentions, such as abbreviation pairs (e.g., “Estrogen receptor (ER) alpha”), include more than one reference to the same concept. These therefore appear very similar to composite mentions, complicating the normalization of this type of mention to a specific database identifier. To make this model more robust and able to handle more types of term variation, this study considers abbreviation pairs as a type of composite mention. After simplifying the mentions, it is easier to map the resulting concept mentions to a database or controlled vocabulary identifier. In this study, composite mentions can be divided into 5 distinct types (including abbreviation pair) and a mixed type of mentions.

1. Mention with coordination ellipsis: the concepts in this type of mentions share part of the mention region, such as the token “SMAD” in the composite mention “SMADs 1, 5, and 8”, which is shared to “SMAD1”, “SMAD5”, “SMAD8”.

2. Range mention: Like mentions with coordination ellipsis, these mentions share part of the mention region, however this type represents a range of entities rather than a discrete set. For example, the range “SMAD 2 to 4” can be decomposed to “SMAD2”, “SMAD3”, “SMAD4”.

3. Individual mention: this is an independent composite mention. All concepts can be separated into non-overlapping spans (e.g., “BTK/ITK/TEC/TXK” can be decomposed to “BTK”, “ITK”, “TEC” and “TXK”).

4. Overlap abbreviation pair mention: The long form and short form share some tokens, like “COUP (chicken ovalbumin upstream promoter) transcription factor” which can be decomposed to “COUP transcription factor” and “chicken ovalbumin upstream promoter transcription factor”. But the two concepts indicate the same database identifier.

5. Individual abbreviation pair mention: this is an independent composite mention. The same to overlap abbreviation pair mention, the two concepts indicate the same database identifier. For example, “ectodermal dysplasia (EDA)” can be decomposed to “ectodermal dysplasia” and “EDA”.

6. Mixed mention: It is a mixed mention of any two above types, like “high mobility group (HMG) protein 1 and 2” which can be decomposed to “high mobility group protein 1”, “HMG protein 1”, “high mobility group protein 2” and “HMG protein 2”.

The 3rd and 4th types of mentions do not overlap and therefore should be identified as separate by the mention recognition model. However, since these are easily confused with composite mentions, the mention recognition model may identify the wrong boundaries. It is therefore important for the mention simplification model to also identify these correctly and separate them into individual concepts. However, not all mentions which follow these patterns should be separated. For example, “ubiquitin-activating enzyme (E1)” and “metalloprotease/disintegrin/cysteine-rich protein 9” both represent single concepts and should not be separated. These cases might not be easy for pattern identification method. Therefore, a robust method should be able to deal with these exceptions.

In this study, we developed SimConcept, a simplification method for composite mentions, which aims to help concept normalization in the end. It detects various concepts of a composite mention. Here are the three main contributions: 1) SimConcept can handle six types of composite mentions, more than any other methods previously reported; 2) When applied to the three bio-concepts (i.e., gene, disease, and chemical), our method achieved state-of-the-art performance and 3) Based on our success on more than one entity type, our approach is shown to be robust and generalizable.

2. METHODS

Our method consists of two modules as shown in Figure 1. The first module consists of a conditional random field model. In this module, the input mention is separated into tokens and each token assigned labels according to the most likely sequence of states through the model. The second module reassembles the tokens into individual mentions using a pattern identification method.

![Figure 1. Architecture of the SimConcept mention simplification method](image)

2.1 Conditional random fields model

To recognize the composite mentions, we observed the composition of those mentions and defined nine states for building conditional random fields (CRF) model [39]: Antecedent (A); Strain/Suffix(S); Conjunction of mentions with coordination ellipse (C); Conjunction of range mentions (C); Left parentheses of abbreviation pair (L); Right parentheses of abbreviation (R);
Right parentheses of abbreviation, but the abbreviation and long form cannot be separated (Rd); Conjunction of individual mentions (I); Redundant (O). The states “C”, “Ca”, “Li”, “R”, “Rd” (L and R/Rd occur in pairs), and “I” are conjunction states which can use to recognize the mention types. If one mention includes two or more conjunction states, this mention would be identified as a mixed mention.

As mentioned above, we regarded this mention simplification problem as a sequence labeling task, as illustrated in a simple example in Table 1. Our implementation uses a linear chain Conditional Random Fields (CRF) [39] provided by CRF++ (http://crfpp.googlecode.com/svn/trunk/doc/index.html). The CRF model defines the conditional probability distribution \( p(Y|X) \) of label sequence \( Y \) given observation sequence \( X \). The lengths of random variable sequences \( X (x_1, ..., x_n) \) and \( Y (y_1, ..., y_n) \) are the same.

\[
p(Y|X) = \frac{\exp(F(X, Y))}{\sum_{Y'} \exp(F(X, Y'))}
\]

The CRF model on \( (X, Y) \) is specified by a vector \( F \) of global features and a corresponding weight vector \( \lambda (\lambda_1, ..., \lambda_i) \).

\[
F(X, Y) = \sum_{i=1}^{n} \lambda_i f_i(y_i, y_{i-1}, X)
\]

is a global feature vector for label sequence \( Y \) and observation sequence \( X \). In our research, the \( y_1, ..., y_n \) indicate the label for the corresponding tokens. To identify the antecedent and conjuncts part of mention.

<table>
<thead>
<tr>
<th>X</th>
<th>BRCA</th>
<th>1 / 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>A</td>
<td>S</td>
</tr>
</tbody>
</table>

The weight \( \lambda \) presents the importance of the feature \( f_i(y_j, y_{j-1}, X) \) and can be obtained from the training data. CRF++ applies L-BFGS [40] which is a quasi-newton algorithm for large scale numerical optimization problem.

### 2.2 CRF Features

We adapted tmVar [41], our previous study on mutation recognition, to this task. We used tmVar’s tokenization and part of its features in SimConcept development. Like tmVar, our tokenization separates uppercase characters, lowercase characters and digits. For example, “SMADs 2 to 4” is separated to “SMAD’, “s”, “2”, “to” and “4”. We adapted tmVar’s features to reflect the difference in input between tmVar (i.e., documents) and SimConcept (i.e., individual mentions). After reviewing the evidence for different token types of a mention, we defined several suffixes, prefixes and some semantic types for identifying bioconcepts (i.e., gene, disease, and chemical) mention characteristics. In particular, most mention suffixes for disease and chemical mentions are not digits, for example “breast and ovarian cancer” (disease) and “b-sitosteryl and stigmasteryl linoleates” (chemical), which might be difficult to recognize without any semantic evidence. Therefore, we collected the semantic features used in some previous studies [41-43] and grouped the suffixes PREFIXES we defined into semantic feature types such as those shown below.

- **Chemical Suffix**: yl, ylidyne, oyl, sulfony, one, ol, carboxylic, amide, ate, acid, ium, ylium and etc.
- **Chemical Alkane Stem**: meth, eth, prop, tetracos
- **Chemical Trivial Ring**: benzene, pyridine, toluen
- **Chemical Simple Multiplier**: di, tri, tetra and etc.
- **Disease Suffix**: cancer, helium, lithium, beryllium, boron, carbon and etc.
- **Gene/Protein Suffix**: gene, protein, receptor, factor, element, unit and etc.
- **Family, Complex**: family, subfamily, superfamily, complex

We also continue to use three of tmVar’s features types (i.e., Character features, Case pattern features and Contextual features). Character features include number of digits, number of uppercase and lowercase letters, number of all characters, and specific characters (; , . ?). Case pattern features are created by replacing uppercase alphabetic character to “A” and any lower case to ‘a’. Likewise any number (0-9) is replaced by ‘0’. Moreover, we also merged consecutive letters and numbers to generate additional features, such as “AAA” to “A”. In order to take advantage of contextual information, for a given token we included the token and semantic features of 3 neighboring tokens from each side.

### 2.3 Token reassembly through pattern identification

By observing the characteristics of composite mentions in our training data, we manually defined four patterns to model the six types of composite bioconcept mentions, as shown in Figure 2. To simplify mentions, we distinguish between the antecedent region (green), conjuncts region (frame), conjunct candidate (blue) and conjunctions (red). The tokens in antecedent region should be present in all possible mentions. The tokens in conjuncts region should be replaced by all possible conjunct candidates in this region. Every conjuncts region consists of at least one conjunct. Conjunctions are used to separate individual conjunct candidates.

In our definition, every mention can map to one of the patterns. Range mentions and mentions with coordination ellipsis map to Pattern 1. As shown in Figure 3a, the “ORP-2 to -4” is a range mention which can be separated to “ORP-2” (antecedent region), and “2 to -4” (conjuncts region). In conjuncts region, all possible candidates (i.e., 2, 3 and 4 in “2 to -4”) belong to one of the possible mentions. Therefore, “ORP-2 to -4” is reassembled to “ORP-2”, “ORP-3” and “ORP-4”. In another similar case, the “ORP-1 and -2” is similar to “ORP-2 to -4”. The major difference is the conjunction (i.e., “and”). In this case, “-1” and “-2” in conjuncts region are independent. Therefore “ORP-1 and -2” becomes “ORP-1” and “ORP-2”.

---

*References and further details are provided in the original document.*
Figure 2. Patterns for formulating bioconcept mentions. Note that in Patterns 1 & 2, it is common to have Antecedent appearing either at the beginning or the end but not both locations (i.e. component a and d may not appear together in one mention).

Pattern 1:

Ex. ORP-2 to 4

Pattern 2:

Ex: human neurokinin-3 (NK-3) receptor

Pattern 3:

Ex: ORP-2 and ORP-2

Pattern 4:

Ex: estrogen receptor 1 (ESR1)

Figure 3. The strategy of reassembly for mention with coordination ellipsis, range mention and abbreviation

Observing these two cases, it becomes clear that the difference between range mentions and mentions with coordination ellipsis is the conjunctions. In case the conjunction is recognized as a conjunction of range mentions (C state), all values in the range of these two suffixes should be considered as conjunct candidates. Otherwise, once the conjunction is recognized as a conjunction of mentions with coordination ellipsis (C state), the candidates in the conjuncts region are independent to each other, and dependent to the antecedent region. Therefore, the reassembly mentions are the combinations of antecedent region and each conjunct candidate.
As shown in Figure 3b, individual and overlap abbreviation pair mention belong to Pattern 2. The pair (long form “neurokinin-3" and abbreviation “NK-3") of abbreviation mentions is in the conjuncts region. Therefore, the two candidates, long form and abbreviation, are reassembled with antecedent region individually. We detected the long form region by applying the Ab3P abbreviation identification tool [44]. Thus, we are able to identify the conjuncts region in these mentions. Patterns 3 and 4 in Figure 2 are relatively easier than Pattern 1 and 2. Since the patterns do not contain a conjuncts region, assembling the individual mentions only requires splitting conjunctions and parentheses. As shown in Figure 3c/d, the mentions can be separated individually.

In addition to the above types, mixed mention is more complicated. We defined a two phase strategy to divide concepts. In the 1st phase we split the mention using Patterns 3 or 4, which do not contain any conjuncts region. In this phase, all conjunctions of range (Cn), mention with coordination ellipsis (C) and abbreviation parentheses (L/Rn) are considered as part of antecedent region. In the 2nd phase, the mention is decomposed by Pattern 1 or 2. In this phase we start to face the conjuncts region. As shown in Figure 4, “interferon gamma (IFN-gamma)-inducible protein 10 (gamma IP-10)" is split to “interferon gamma (IFN-gamma)-inducible protein 10" and “gamma IP-10" by cutting in 1st phase. According the states L/Rn, which have been identified in “interferon gamma (IFN-gamma)-inducible protein 10", the 2nd phase should choose Pattern 2 for simplifying. We therefore obtained “interferon gamma-inducible protein 10" and “IFN-gamma-inducible protein 10" from the 2nd phase simplification.

In other words, the main idea of this two phase strategy is to retain all sub-mentions with a conjuncts region in the second phase. Since the sub-mentions which map to Patterns 1 & 2 are more complicated and cannot be separated individually, those sub-mentions will be processed in the second step.

### 2.4 The SimConcept Corpus

The SimConcept corpus was compiled using 5 datasets: three for genes, one for diseases and one for chemicals. For genes, we integrated the BioCreative II gene normalization task training (281 abstracts) and test (262 abstracts) corpora and the 151 GIA test collection (http://it.nlm.nih.gov/DataSets/index.shtml#GIA). In addition, we also collected disease mention corpus from NCBI Disease corpus [18, 45] with 793 abstracts, and sampled Chemical mention corpus from BioCreative IV CHEMDNER task [46] training dataset for 937 abstracts. As shown in Table 2, we collected 2,424 abstracts in total. For each article, in addition to the annotations of all described bioconcept mentions, we appended following annotations: 1) the decomposed mentions, such as “BRCA1" and “BRCA2" of “BRCA1/2“; 2) the five types of composite mentions (e.g., “mention with coordination ellipsis“); 3) the states of tokens (“BRCA 1 / 2 “). We used PubMed [47-49], a web-based annotation tool to annotate the corpus. The distributions of the five composite mention types (Cx: Range mention, C: mention of coordination ellipsis, I: individual mentions, IA: individual abbreviation, and OA: overlap abbreviation,) in Table 2 are different between the three sets. Chemicals contain significantly more range mentions than either disease or genes, and diseases contain more individual abbreviations than chemicals or genes. The distribution for genes is more even across all types than either diseases or chemicals.

**Table 2. Descriptive statistics for the SimConcept corpus.**

The numbers of composite mentions (of different types) are first listed followed by the numbers of individual mentions after decomposition in parentheses.

<table>
<thead>
<tr>
<th>Concept</th>
<th># of abstracts</th>
<th>Five types of composite mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>694</td>
<td>810 (1895) 14 (60) 101 (246) 442 (1089) 253 (534) 41 (107)</td>
</tr>
<tr>
<td>Disease</td>
<td>793</td>
<td>1012 (2293) 2 (18) 245 (583) 303 (809) 486 (1045) 52 (123)</td>
</tr>
<tr>
<td>Chemical</td>
<td>937</td>
<td>1012 (2944) 99 (505) 201 (771) 496 (1389) 302 (716) 0 (0)</td>
</tr>
</tbody>
</table>

**3. EXPERIMENTAL RESULTS AND DISCUSSION**

To evaluate our method, we used leave-one-out cross validation on the three sets (i.e., gene, disease and chemical). Table 3 shows the results of our evaluation, where we see that the overall performance is high for all three entity types.

**Table 3. The Statistic of SimConcept corpus**

<table>
<thead>
<tr>
<th>Concept</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>88.41%</td>
<td>90.19%</td>
<td>89.29%</td>
</tr>
<tr>
<td>Disease</td>
<td>87.03%</td>
<td>84.07%</td>
<td>85.52%</td>
</tr>
<tr>
<td>Chemical</td>
<td>85.36%</td>
<td>82.78%</td>
<td>84.04%</td>
</tr>
</tbody>
</table>

The chemical corpus includes two types of mentions which are not addressed by our patterns. The first exception is a joint
mention which the second mention uses coreference to indicate the previous mention, such as “3-0-propargylated betulinic acid and its 1,2,3-triazoles”. The pronoun “its” represents the previous mention “3-0-propargylated betulinic acid”. Therefore, this composite mention contains two individual mentions “3-0-propargylated betulinic acid” and “1,2,3-triazoles of 3-0-propargylated betulinic acid”. The other exception mention is a variant of the continuous type. For example, “tenofovir mono- or diphosphate”, “phosphate” is the postcedent region. Even though our applied tokenization splits on special characters, digits, lowercase and uppercase letters, which is more specific than general tokenization, it is still unable to split the conjuncts region and postcedent region. We have ignored these two types of mentions since they are rare.

To assess the performance on each composite mention type, we computed results shown in Table 4. There are only two range mentions in the disease set, and we therefore ignored these. There are also no overlap mentions in the chemical set. Since two exception mentions belong to continuous mention type in chemical corpus, the performance of continuous mention becomes lower.

Table 4. The evaluation of individual mention types. Scores are f-measures.

<table>
<thead>
<tr>
<th></th>
<th>Gene</th>
<th>Disease</th>
<th>Chemical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual abbreviation</td>
<td>91.11%</td>
<td>84.21%</td>
<td>84.15%</td>
</tr>
<tr>
<td>Overlap abbreviation</td>
<td>80.92%</td>
<td>91.53%</td>
<td>N/A</td>
</tr>
<tr>
<td>Mention with coordination ellipsis</td>
<td>75.88%</td>
<td>77.46%</td>
<td>61.96%</td>
</tr>
<tr>
<td>Range mention</td>
<td>91.34%</td>
<td>N/A</td>
<td>94.14%</td>
</tr>
<tr>
<td>Individual mention</td>
<td>90.24%</td>
<td>87.13%</td>
<td>87.34%</td>
</tr>
<tr>
<td>Mixed mention</td>
<td>81.64%</td>
<td>81.25%</td>
<td>83.76%</td>
</tr>
<tr>
<td>All composite mentions</td>
<td>89.29%</td>
<td>85.88%</td>
<td>84.04%</td>
</tr>
</tbody>
</table>

Table 5. The SimConcept contribution on gene normalization performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenNorm + SimConcept</td>
<td>87.01%</td>
<td>86.13%</td>
<td>86.57%</td>
</tr>
<tr>
<td>GenNorm + heuristic rules</td>
<td>86.78%</td>
<td>85.23%</td>
<td>86.00%</td>
</tr>
<tr>
<td>GenNorm</td>
<td>86.72%</td>
<td>84.09%</td>
<td>85.38%</td>
</tr>
</tbody>
</table>

Table 6. The SimConcept contribution on disease normalization performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNorm + SimConcept</td>
<td>80.91%</td>
<td>79.23%</td>
<td>80.06%</td>
</tr>
<tr>
<td>DNorm</td>
<td>80.69%</td>
<td>76.85%</td>
<td>78.72%</td>
</tr>
</tbody>
</table>

As mentioned in introduction, this study is aimed for helping bioconcept normalization. We therefore applied SimConcept in GenNorm [21] and DNorm [18], and evaluated on the test sets of BioCreative II gene normalization task [12] and NCBI disease corpus [50], respectively (no normalized chemical corpus is available). To avoid training on the test set, the training set for SimConcept excluded the test corpora for GenNorm and DNorm. As shown in Table 5 and Table 6, using SimConcept can further improve the state-of-the-art performance for 1.17% in F-measure (P-value=0.02) for gene normalization and 1.34% in F-measure (P-value=0.03) for disease normalization. We also applied the heuristic rules used in previous gene normalization studies [51, 52] and showed the result in second row of Table 5. Our set of heuristics includes 9 rules. Those rules are defined by regular expressions to recognize the conjuncts at the end of the mention (e.g., detecting “1” and “2” in “BRCA1/2”) and handle some mentions containing coordination ellipses and ranges. However, the composite mentions which are not considered in the refinement of heuristic rules cannot be recognized. This comparison shows that using heuristic rules is not as robust as SimConcept. As also shown in Table 5, using heuristic rules raises performance about half as much as SimConcept.

In order to examine the contribution of individual feature types, we performed a feature ablation study where different feature types were removed from the entire set of features one at a time. As shown in Table 7, the largest drop in performance was due to the removal of token features, followed by semantic and character features. The removal of case pattern or contextual features had little effect on final performance. In addition to removing features, we also changed the order of CRF model from order 2 to order 1. The result shows order 2 performs better than order 1.

Table 7. Performance decrease when removing features. (Evaluated on gene corpus)

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimConcept</td>
<td>88.41%</td>
<td>90.19%</td>
<td>89.29%</td>
</tr>
<tr>
<td>- Token features</td>
<td>86.94%</td>
<td>88.50%</td>
<td>87.71%</td>
</tr>
<tr>
<td>- Semantic features</td>
<td>87.90%</td>
<td>89.29%</td>
<td>88.58%</td>
</tr>
<tr>
<td>- Character features</td>
<td>87.78%</td>
<td>89.87%</td>
<td>88.81%</td>
</tr>
<tr>
<td>- Pattern features</td>
<td>88.30%</td>
<td>90.10%</td>
<td>89.19%</td>
</tr>
<tr>
<td>Order 2 ➔ Order 1</td>
<td>86.52%</td>
<td>88.39%</td>
<td>87.44%</td>
</tr>
</tbody>
</table>

4. CONCLUSION

In this study, we present SimConcept – a method to handle the task of composite named entity simplification. We integrated a CRF-based method with a pattern identification strategy to systematically decompose the six types of composite mentions. To handle the three most fundamental bioconcepts, we re-annotated the composite mentions in five existing corpora for gene (BioCreative 2 GN task/train/test corpus and NLM GIA corpus), disease (NCBI disease corpus) and chemicals (BioCreative IV ChemDNER task corpus), and used these to evaluate SimConcept. The results show that SimConcept handles composite mention simplification issue effectively.

We further used SimConcept to assist the bioconcept normalization task. The result suggests that SimConcept is helpful for improving normalization performance. Our approach should generalize to other entity types in addition to the three concepts that were the focus of this study: genes, diseases and chemicals.

5. ACKNOWLEDGMENTS

Funding: This research was supported by the NIH Intramural Research Program, National Library of Medicine.

6. REFERENCES


proceedings of BioCreative (Washington DC, USA, April 5, 2012). BioCreative 2012 workshop


